STATISTICAL IDENTIFICATION OF METABOLIC REACTIONS CATALYZED BY GENE PRODUCTS OF UNKNOWN FUNCTION

by

LIANQING ZHENG

M.S., Western Michigan University, 2006

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2013

Abstract

High-throughput metabolite analysis is an approach used by biologists seeking to identify the functions of genes. A mutation in a gene encoding an enzyme is expected to alter the level of the metabolites which serve as the enzyme's reactant(s) (also known as substrate) and product(s). To find the function of a mutated gene, metabolite data from a wild-type organism and a mutant are compared and candidate reactants and products are identified. The screening principle is that the concentration of reactants will be higher and the concentration of products will be lower in the mutant than in wild type. This is because the mutation reduces the reaction between the reactant and the product in the mutant organism.

Based upon this principle, we suggest a method to screen the possible lipid reactant and product pairs related to a mutation affecting an unknown reaction. Some numerical facts are given for the treatment means for the lipid pairs in each treatment group, and relations between the means are found for the paired lipids. A set of statistics from the relations between the means of the lipid pairs is derived. Reactant and product lipid pairs associated with specific mutations are used to assess the results.

We have explored four methods using the test statistics to obtain a list of potential reactant-product pairs affected by the mutation. The first method uses the parametric bootstrap to obtain an empirical null distribution of the test statistic and a technique to identify a family of distributions and corresponding parameter estimates for modeling the null distribution. The second method uses a mixture of normal distributions to model the empirical bootstrap null. The third method uses a normal mixture model with multiple components to model the entire distribution of test statistics from all pairs of lipids. The argument is made that, for some cases, one of the model components is that for lipid pairs affected by the mutation while the other components model the null distribution. The fourth method uses a two-way ANOVA model with an interaction term to find the relations between the mean concentrations and the role of a lipid as a reactant or product in a specific lipid pair. The goal of all methods is to identify a list of findings by false discovery techniques. Finally a simulation technique is proposed to evaluate properties of statistical methods for identifying candidate reactant-product pairs.

STATISTICAL IDENTIFICATION OF METABOLIC REACTIONS CATALYZED BY GENE PRODUCTS OF UNKNOWN FUNCTION

by

LIANQING ZHENG

M.S., Western Michigan University, 2006

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2013

Approved by:

Major Professor
Gary L. Gadbury

Copyright

Abstract

High-throughput metabolite analysis is an approach used by biologists seeking to identify the functions of genes. A mutation in a gene encoding an enzyme is expected to alter the level of the metabolites which serve as the enzyme's reactant(s) (also known as substrate) and product(s). To find the function of a mutated gene, metabolite data from a wild-type organism and a mutant are compared and candidate reactants and products are identified. The screening principle is that the concentration of reactants will be higher and the concentration of products will be lower in the mutant than in wild type. This is because the mutation reduces the reaction between the reactant and the product in the mutant organism.

Based upon this principle, we suggest a method to screen the possible lipid reactant and product pairs related to a mutation affecting an unknown reaction. Some numerical facts are given for the treatment means for the lipid pairs in each treatment group, and relations between the means are found for the paired lipids. A set of statistics from the relations between the means of the lipid pairs is derived. Reactant and product lipid pairs associated with specific mutations are used to assess the results.

We have explored four methods using the test statistics to obtain a list of potential reactant-product pairs affected by the mutation. The first method uses the parametric bootstrap to obtain an empirical null distribution of the test statistic and a technique to identify a family of distributions and corresponding parameter estimates for modeling the null distribution. The second method uses a mixture of normal distributions to model the empirical bootstrap null. The third method uses a normal mixture model with multiple components to model the entire distribution of test statistics from all pairs of lipids. The argument is made that, for some cases, one of the model components is that for lipid pairs affected by the mutation while the other components model the null distribution. The fourth method uses a two-way ANOVA model with an interaction term to find the relations between the mean concentrations and the role of a lipid as a reactant or product in a specific lipid pair. The goal of all methods is to identify a list of findings by false discovery techniques. Finally a simulation technique is proposed to evaluate properties of statistical methods for identifying candidate reactant-product pairs.

Table of Contents

List of Figures ix
List of Tables xii
Acknowledgements xiv
Chapter 1 - Introduction 1
1.1. The Functional Genomics Tools 1
1.2. What is Metabolomics? 3
1.3. Metabolite / lipid Data Analysis 5
1.3.1. The Metabolome/Lipidome Data Analysis Work Flow 5
1.3.2. The Lipidomics Pathways 7
1.3.3. The Fundamental Scheme Used in the Pathway Data Analysis 8
1.4. Dissertation Outline 10
Chapter 2 - Literature Review 12
2.1. Pathway Analysis Overview 12
2.2. FANCY Approach 15
2.3. Correlation Analysis 18
2.3.1. Introduction to Correlation Analysis 18
2.3.2. Roots-aerial Datasets Correlation Analysis in Fukushima et al. (2011) 19
2.3.3. Results for Roots-Aerial Datasets 21
2.4. False Discovery Rates and Mixture Model Approaches in High-dimensional Data Analysis 25
Chapter 3 - Exploratory Data Analysis 28
3.1. Difficulties in Analyzing the Lipid Pathway Experimental Data 28
3.2. Data Manipulation 29
3.2.1 Centering and Scaling 29
3.2.2 Using Variable y to Screen for the Population Lipid Pairs of Interest 34
3.3. Illustration of Treatment Mean Relationships from Scatter Plots 35
3.3.1. Patterns in Scatter Plots When y =2 and y = 1 35
3.3.2. Using Biologically AB Pairs as a Criterion for Identifying all AB Pairs 36
3.4. Can Correlation Analysis Work in our Lipid Experiments? 38

3.5. Some Numerical Facts from the Exploratory Data Analysis 41
3.6. Three Summary Statistics 43
Chapter 4 - A Parametric Bootstrap Null Distribution 47
4.1. Bootstrap Algorithm 47
4.2 Bootstrap Null Distribution 50
4.2.1. Choose Distribution Class for the Null Distribution 50
4.2.2. Determine the Final Parametric Null Distributions and Assess the Goodness-of-fit 53
4.3. Results in *fad2* Dataset 54
4.3.1. Choose the Null Distributions 54
4.3.2. Assess the Final Selected Parametric Null Distributions using the Empirical Distributions and Fitting Results to the Data 56
Chapter 5 - A Mixture Normal Bootstrap Null Distribution 62
5.1. Introduction to the Mixture Normal Distribution 62
5.2. Mixture Normal Bootstrap Null Distribution (MNBN) 62
5.3. Randomization Test for the Treatment Effects 63
5.4. MNBN in fad2 and fad4 Datasets 65
5.4.1. Find the MNBN Distributions 65
5.4.2. The Results from fad4 and fad2 69
Chapter 6 - Bootstrap Methods Under the Equal Mean Hypothesis 72
6.1. Bootstrap Algorithm Under Equal Mean Hypothesis 72
6.2. Parametric Bootstrap Null (PBN) Distribution Fitting under $\mu_F = \mu_G$ 75
6.2.1. The Results from fad2 Dataset Using PBN 75
6.2.2. The Results from fad4 Dataset Using PBN 78
6.3. Mixture Normal Bootstrap Null (MNBN) distribution Fitting Under the Equal Mean Hypothesis 80
6.3.1. The Results from fad2 Dataset Using MNBN 80
6.3.2. The Results from fad4 Dataset Using MNBN 83
6.4. Discussion 85
Chapter 7 - A Mixture Model to Fit the R_T Distribution in the Data 87
7.1. Normal Mixture Distributions for R_T 87
7.2. Test the Number of Components in the Normal Mixture Models 90

7.3. Estimation and Confidence Intervals for the Parameters in Normal Mixture Models 93
7.4 The Results for fad2 Dataset by Using Three-component Normal Mixture..... 95
Chapter 8 - ANOVA Approach for the Pathway Analysis..... 100
8.1. Two-way ANOVA Model with an Interaction Term 100
8.2. ANOVA Interaction Test Results for the 9 Lipid Datasets 101
8.3. ANOVA Interaction Test Results for the Roots-aerial data 103
Chapter 9 - Simulation of Realistic Data 105
9.1. Introduction..... 105
9.2. Data Simulation Algorithms 105
9.3. Simulation Using fad4 Dataset 110
9.3.1. The Characteristics of the Simulated Datasets..... 111
9.3.2. The MNBN Method to Fit to the Bootstrap Distributions..... 114
9.4. Results and Discussion 115
Chapter 10 - Summary and Future Work..... 118
10.1 Summary of This Dissertation 118
10.2 Future Direction..... 119
10.2.1. Intersection-Union Test 119
10.2.2. Dependence in the Data 120
10.2.3. Variance Structure in the Lipid Pairs..... 120
References..... 122
Appendix A - The Lipidomics Experiment Information 129
Appendix B – R Programs..... 131
B.1: Correlation Analysis from Fukushima et. al. (2011) in Chapter 2..... 131
B.2: Produce Venn Diagram in Figure 2.4. 133
B.3: Produce Test Statistics SSD in Chapter 3. 135
B.4: Generate Bootstrap Under F = G and Make Plots in Chapter 4..... 136
B.5: Mixture Normal Distributions in Chapter 5. 143
B.6: Mixture Normal Distributions to Fit the Empirical Distribution of R_T in Chapter 7..... 145
B.7: Simulation in Chapter 9. 149

List of Figures

Figure 1.1: The information flow chart in an organism (Hollywood et al. 2006) 1
Figure 1.2: The structure of 18:1, oleic acid 4
Figure 1.3: The lipid data analysis work flow chart (Welti and Wang, 2004) 6
Figure 1.4: A schematic drawing of the lipid pathway network 7
Figure 1.5: The scheme used to find reactant and product AB lipid pairs 9
Figure 1.6: The concentration level relationships of reactant and product AB lipid pairs 9
Figure 2.1: Comparison of some main methods for detecting a metabolite pathway 13
Figure 2.2: Plot of the co-response θ as a function of Ω 16
Figure 2.3: Correlation analysis results for comparing the 3 genotypes in *Arabidopsis* 22
Figure 2.4: The numbers of common correlations from the three genotypes 23
Figure 2.5: Illustration of the correlation analysis using slopes in each treatment group 24
Figure 3.1: Example scatter plots of one lipid pair before and after scaling 34
Figure 3.2: Four representative scatter plots from $y = 2$ and $y = 1$ 35
Figure 3.3: Scatter plots of some biologically functional lipid pairs in *fad2* 37
Figure 3.4: A typical reactant and product candidate pair after scaling. 38
Figure 3.5: The number of significant correlations comparison in all 9 datasets 41
Figure 3.6: The relations of the treatment means for an AB lipid pair 42
Figure 3.7: The scatter plot and the treatment mean plot when $y = 2$ from dataset fad2 43
Figure 3.8: The distributions of the test statistics. 44
Figure 3.9: Illustration of R statistic from the scatter plot using tg and SSD in dataset fad2 45
Figure 4.1: Schematic of the bootstrap procedure for generating the three test statistics from each bootstrap sample 48
Figure 4.2: Flow chart for generating one null distribution of the statistics using the maximum likelihood estimation procedure in the b^{th} bootstrap sample. 51
Figure 4.3: The empirical distributions of the three test statistics from 200 bootstraps 54
Figure 4.4: The distribution of K-S test statistic D for all three statistics tg^*, SSD^* and R_T^* 56
Figure 4.5: ECDFs matches with the selected null Weibull CDFs for the statistic SSD 58

Figure 4.6: tg statistic parametric bootstrap null distribution overlaid with the tg distribution in fad2 59
Figure 4.7: SSD parametric bootstrap null distribution overlaid with the SSD distribution in *fad2* 59
Figure 4.8: R_T bootstrap null distribution overlaid with the distribution in fad2 60
Figure 5.1: Bar charts for the distribution of the p values from the randomization test 65
Figure 5.2: The bootstrap distribution overlaid with the R_T distribution in fad4 data 66
Figure 5.3: Two-component MNBN distribution of R_T^* in fad4 data 67
Figure 5.4: Three-component MNBN distribution of R_T^* in fad4 data 68
Figure 6.1: The R_T^* bootstrap distribution comparison under the assumptions of $F = G$ and $\mu_F = \mu_G$ 74
Figure 6.2: The distribution of K-S test statistic D for all three statistics tg^*, SSD^* and R_T^* in fad2 76
Figure 6.3: The 95th percentile empirical bootstrap distribution overlaid with the 95th percentile Weibull distribution and the real data from fad2 77
Figure 6.4: The distribution of K-S test statistic D for all three statistics tg^*, SSD^* and R_T^* in fad4 79
Figure 6.5: The 95th percentile empirical bootstrap distribution overlaid with the 95th percentile Weibull distribution and the real data from fad4 80
Figure 6.6: The R_T^* bootstrap empirical distribution overlaid with the R_T distribution for fad2 under the hypothesis of $\mu_F = \mu_G$ 81
Figure 6.7: Two-component MNBN distribution of R_T^* in fad2 82
Figure 6.8: The R_T^* bootstrap empirical distribution overlaid with the R_T distribution for fad4 under the assumption of $\mu_F = \mu_G$ 83
Figure 6.9: Two-component MNBN distribution of R_T^* in fad4 data 84
Figure 7.1: The two-component mixture normal distributions fitting to all the fads datasets 90
Figure 7.2: The three-component mixture normal distributions fitting to all the fads datasets 91
Figure 7.3: The mixture model fit of R_T in fad2 with three normal components 94
Figure 7.4: The posterior probabilities of the R_T for each lipid pair in fad2 96

Figure 7.5: The proportion of significant pairs and the proportion of significant distinct reactants in fad2 98
Figure 8.1: Interaction plots for a lipid pair with a significant interaction 102
Figure 8.2: Comparison of the number of significant pairs in all 9 datasets 103
Figure 8.3: Comparison of significant metabolite pairs from the *roots-aerial* data using interaction p values 104
Figure 9.1: The scatter plot tg versus SSD and the R_T distribution in the actual dataset fad4 ... 111
Figure 9.2: The scatter plot of tg versus SSD from the four simulations 112
Figure 9.3: The R_T distribution in the four simulations 113
Figure 9.4: Normal mixture model fit to the actual fad4 dataset 114
Figure 9.5: Two-component normal mixture model fit in the four simulations 115
Figure 10.1: The distribution of statistic F from the equal variance test 121

List of Tables

Table 1.1: Abbreviations used in this dissertation (Welti and Wang, 2004) 4
Table 1.2: Some techniques used in metabolic studies (Dunn and Ellis, 2005; Hall, 2005; Fiehn, 2006; Nielse et al. 2005) 6
Table 2.1: Example results of the FANCY method shown with the angle variable θ 17
Table 2.2: The sample sizes in roots-aerial datasets 19
Table 2.3: The sample correlations in roots-aerial datasets 20
Table 2.4: Number of significant correlations in each tissue and genotype/treatments 21
Table 2.5: Hypotheses testing terms for FDR in the high-dimensional data analysis. 26
Table 3.1: Partial data structure from fad2 30
Table 3.2: Illustration of the centered and scaled data 33
Table 3.3: Number of significant Spearman's correlations in the 9 experiments 39
Table 3.4: Number of significant Pearson's correlations in the 9 experiments 40
Table 3.5: Matching number between the biologically functional lipid pairs and the pairs with significant correlation differences from Pearson's correlation analysis 40
Table 4.1: The counts of the minimum K-S test statistic D_{Min} for the 4 candidate distributions.. 55
Table 4.2: The final null distributions of the three statistics from the chosen parametric distribution with the parameter of estimates 56
Table 5.1: Number of the smallest p values, number of p values, and proportion of the smallest p values in each dataset 64
Table 5.2: The MLEs and the log-likelihood value for mixture normal null distribution in fad4 with two components and five parameters $\pi_1,\ \mu_1,\ \sigma_1\ ,\ \mu_2,\ \sigma_2$. 67
Table 5.3: The MLEs and the log-likelihood value for mixture normal null distribution in fad4 with three components and seven parameters $\pi_1, \pi_2\ ,\ \mu_1,\ \sigma_1\ ,\ \mu_2,\ \sigma_2\ ,\ \mu_3,\ \sigma_3$. 68
Table 5.4: The list of significant findings using the MNBN distribution under the null hypothesis F = G 70
Table 6.1: The means and the five number summary of the R_T^* bootstrap sample distributions in fad2 and fad4 datasets under the two assumptions $F = G$ and $\mu_F = \mu_G$. 74

Table 6.2: The counts of the minimum K-S test statistic D_{Min} for the 4 candidate distributions in fad2 dataset. 75
Table 6.3: The final null distributions of the three statistics from the chosen parametric distribution with the parameter of estimates 77
Table 6.4: The counts of the minimum K-S test statistic D_{Min} for the 4 candidate distributions in fad4 under the assumption of $\mu_F = \mu_G$ 78
Table 6.5: The final null distributions of the three statistics from the chosen parametric distribution with the parameter of estimates in fad4 79
Table 6.6: The MLEs and the log-likelihood value for mixture normal null distribution in fad2 with two components and five parameters: π_1, μ_1, σ_1, μ_2, and σ_2. 81
Table 6.7:The MLEs and the log-likelihood value for mixture normal null distribution in fad4 with two components and five parameters π_1, μ_1, σ_1, μ_2, σ_2. 84
Table 6.8: Comparison of the results using PBN and MNBN methods under the null hypothesis $\mu_F = \mu_G$ and $F = G$ 85
Table 7.1: The confidence intervals for the parameters in a normal mixture with three components including eight parameters π_1, π_2, μ_1, μ_2, μ_3, σ_1, σ_2 and σ_3 94
Table 7.2: The results for the significant lipid pairs using 17 biologically functional lipid pairs from three-component mixture model 97
Table 7.3: The 25 distinct reactants from the significant lipid pairs using the lowest posterior probability for the biological pair PC36_2_PC36_3 from the last row of Table 7.4 98
Table 8.1: The top 6 significant lipid pairs from the results in *fad2* data with *lfdr* < 0.001 102
Table 9.1: The results from the 4 simulations using the MNBN method 116

Acknowledgements

First, I would like to thank Dr. John Boyer and Dr. James Neill in the Department of Statistics at Kansas State University for their support and encouragement during my study.

The greatest appreciation goes to my advisor Dr. Gary Gadbury. This research work couldn't be done without his guidance, encouragement and invaluable advice.

I am very grateful to my committee members Dr. Gary Gadbury, Dr. James Neill, Dr. Haiyan Wang and Dr. Ruth Welti. I learned a great deal from all of them. Dr. Neill's comments and advice have set my goal and motivation for the graduate study. Dr. Wang's suggestions have always encouraged me to move forward. I would like to thank Dr. Ruth Welti for serving on my committee and also I want to thank her for preprocessing the datasets for us. I appreciate her constructive comments and useful discussions. This dissertation couldn't be done without Dr. Welti's contribution.

I would like to thank all the faculty members in the Department of Statistics at Kansas State University. Thank you all for teaching me great deal of knowledge from inside and outside of the classroom. A great thanks to Dr. James Higgins for his invaluable advice on teaching, helpful comments and discussions on how to solve problems.

Finally, I want to thank my family, my husband Ximao and my daughters Emily and Ellen for their constant support.

Chapter 1 - Introduction

In the past decade, high dimensional data analysis in genes, metabolites or lipids has been attracting lots of interest not only from biologists but also from statisticians. The unique and high dimensionality characteristics of "omics" data have challenged statisticians to find more suitable techniques to solve biologists' problems. In this research, we will use pathway analysis to find reactant and product lipid pairs which are significantly affected by an unknown function of mutated genes. In this chapter, the relevant biological background is introduced. Then, the treatment conditions and data structure of the experiment are specified, and the statistical scheme used in data analysis is shown. Finally, the outline of this dissertation is given in section 1.4.

1.1. The Functional Genomics Tools

In “omics” high-throughput data analysis, finding gene functions is important to the biologist. The principle that the genetic information flows in the cells of an organism from the gene to protein and to the metabolome has directed biological research for more than five decades. The following flow chart (Hollywood et al. 2006; Holtorf et al. 2002) shows the biological process flow in the organisms:

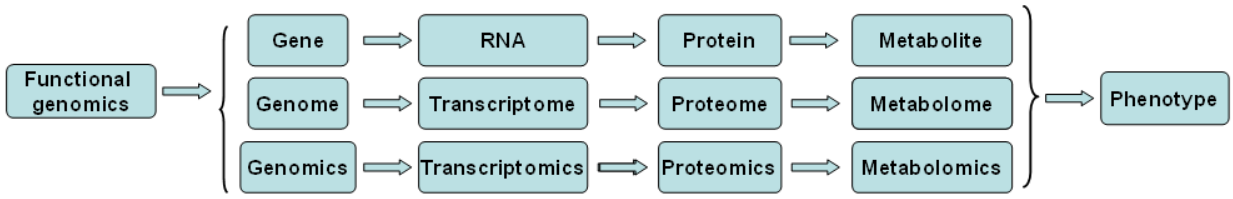

Figure 1.1: The information flow chart in an organism (Hollywood et al. 2006)

In figure 1.1, the first line of the flow chart shows the compound types. The second line shows the global collection of molecules corresponding to the first row. The last row is the broad based approaches to study of the compounds of each type. For example, DNA is the compound type, genome is the collection of DNA, and genomics is the study of all the DNA in an organism (Hollywood et al. 2006). A gene is a nucleotide sequence or sequence of DNA located in chromosomes. It carries genetic information inherited from the parents. In the cell, the genetic information flows from DNA, and passes the genetic code to RNAs. RNAs carry information needed to synthesize proteins. One type of protein is an enzyme which, in the next process, will catalyze reactions among the metabolites. All biological processes are revealed in the phenotype

of the organism. The phenotype reflects the physical characteristics (appearances) of the organisms. Lipids, as one important group of metabolites, play a critical role in the cells' biological processes. The primary functions of lipids include storing energy for the organism, making the membranes of a cell. Biologists sometimes combine all the "omics" (genomics, transcriptomics, proteomics, and metabolomics) to find the gene functions in the living organism.

A functional genomic study as indicated in figure 1.1 can be conducted in two different ways. Biologists can directly study genes by using DNA sequences. There are many DNA sequences now in databases, such as NCBI (The National Center for Biotechnology Information, *http://www.ncbi.nlm.nih.gov/*). Scientists can compare DNA sequences with known sequences from the database by using bioinformatics tools. In this way, they can understand gene functions and/or find new genes.

Biologists can also study genes by analyzing downstream gene products. There are three main strategies that biologists use to derive gene functions through downstream gene products (Allen et al. 2003). The first approach is to use the changes in the transcripts (RNAs) to make an inference about the gene. A popular platform for this approach is to use microarray techniques to study the changes in gene expression. CDNAs, complementary to expressed RNAs, are hybridized to microarray chips, and then one can use the intensities of the probes on a microarray chip to obtain information for genes through statistical analysis. The second approach is to study the expression changes in proteins through high-throughput analysis from protein mass spectrometry or through protein microarray techniques. The third approach is to study the changes in metabolites (such as lipids) using the high-throughput techniques, such as mass spectrometry (MS) (Brügger et al. 1997; Allen et al. 2003). Since the metabolite is the final stage between gene and phenotype, the study of metabolites not only can provide information on gene functions but also can explain the physical characteristics (phenotype) of the organisms.

This dissertation focuses on the third approach by studying the biochemical reaction pathways in the metabolome (lipidome). It is possible that gene sequences can be changed. A change in a sequence is called a mutation (*http://ghr.nlm.nih.gov/handbook/mutationsanddisorders/genemutation*). If the mutation affects any biological process in the cell related to metabolites (metabolism), the mutation effect can be revealed in the compounds of the metabolome. This information provides insight as to the mutated gene (Alberts et al. 2002). In

some plants and animals, the occurrence of a mutation can be controlled. The phenotypes of the mutated organism (mutant) can be compared with those of a non-mutated organism (i.e., wild type) to determine the possible function of the gene that was altered by the mutation. For example, a gene in a plant may be responsible for effectively responding to a stress condition such as drought. These phenotypic changes may or may not be visible.

1.2. What is Metabolomics?

The metabolome is the total collection of the set of small molecule metabolites that the biological process leaves behind (Oliver et al. 1998; Oliver 2002; Griffin and Vidal-Puig, 2008; Dunn et al. 2005). The metabolome includes metabolic intermediates, hormones and some other end products of the metabolism. Unlike the genome and the proteome whose elements are composed of similar building blocks, the metabolome is a group of dynamic molecules with various structures and forms in the biochemical process. Biologists use metabolic profiling to get a "snapshot" of the composition of metabolites to understand biomolecular functions within organisms.

What is lipidomics? One part of the metabolome, the lipidome, plays an important role in the biochemical processes in the cell. Lipids are compounds of biological origin that are poorly soluble in water but are soluble in nonpolar solvents (Blei and Odian 2006). Some lipids are structural components of the cell membranes. Others provide energy for metabolism. Still others serve protective functions; Some are involved in transfer of signals within or among cells and tissues. Lipids can be classified into neutral lipids and polar lipids, or can be classified into complex lipids and simple lipids (Welti and Wang 2004). They include fats, sterols, vitamins, fatty acids, and many others.

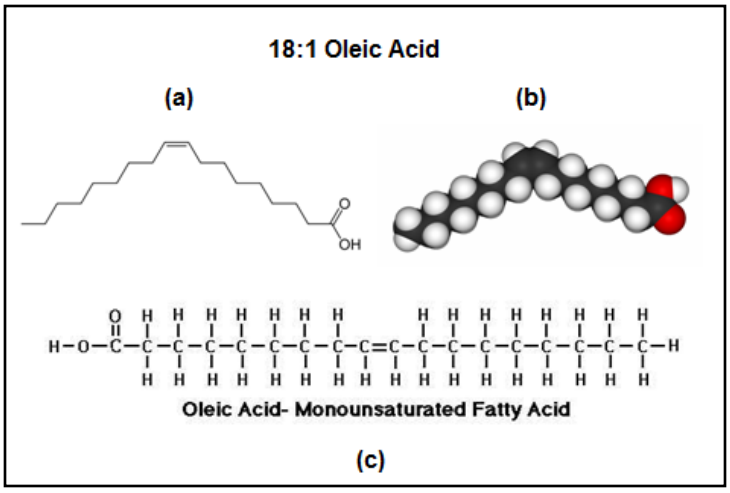

Figure 1.2: The structure of 18:1, oleic acid

(a). The symbol of oleic acid compound with the carboxyl functional group COOH and a long carbon chain tail with a double bound in the middle. **(b)**. Oleic acid model that chemists use to show the structure of a compound. The two red oxygen atoms consist of the carboxyl functional group. **(c)**. Chemical formula to display the biochemical structure of oleic acid. Pictures from: *http://www.raw-milk-facts.com/fatty_acids_T3.html and http://en.wikipedia.org/wiki/Oleic_acid*.

Table 1.1: Abbreviations used in this dissertation (Welti and Wang, 2004)

DGDG	Digalactosyldiacylglycerol
MGDG	Monogalactosyldiacylglycerol
PC	Phosphatidylcholine
LysoPC	Lyso phosphatidylcholine
PE	Phosphatidylethanolamine
PG	Phosphatidylglycerol
LysoPG	Lyso phosphatidylglycerol
PA	Phosphatidic acid
PI	Phosphatidylinositol
PS	Phosphatidylglycerol
sfd	suppressor of fatty acid desaturase deficiency
MS	mass spectrometry
ESI-MS/MS	Electrospray ionization tandem mass spectrometry
fad	fatty acid desaturase6
WT	Wild type organism
MT	Mutant type organism
PCA	Principal component analysis

Since lipids have various chemical structures, I will use one lipid compound oleic acid as an illustration. Figure 1.2 shows the structure of the oleic acid 18:1 lipid. In the lipid compound 18:1, oleic acid, 18 carbon atoms form a long chain with 16 carbon-carbon single bonds and one double bond. The notation 18:1 means an 18 carbon atom chain including only one double bond

in the compound. The two red atoms in the oleic acid model represent oxygen atoms which form the head group of the lipid compound. The double bond commonly exists in the lipid compounds. For example, phosphatidylserine is a lipid compound from the experiment discussed in this dissertation, and the abbreviation is 40:3PS which stands for 40 carbon atoms with 3 double bonds in the lipid. PS represents the head group class. Similar notations are widely used in this project, but with some minor modification. For example, we use the notation PS40_3 to replace 40:3PS. Table 1.1 lists abbreviations for lipid species and other terminology used in this dissertation.

1.3. Metabolite / lipid Data Analysis

Genomics tools that measure genes and gene products cannot always explain the gene's function clearly and precisely (Trethewey, 2001). Increases in levels of RNA and proteins do not necessarily imply that there is more activity in a biological process. A metabolic study is a more useful tool for functional genomics (Bino et al. 2004; Fiehn, 2002). Since the metabolite is the end product in the biochemical process, a more complete picture of the cellular biology, from the functional gene to metabolism of the molecule by the metabolite can be investigated (Wu et al. 2005; Raamsdonk et al. 2001). In addition, plant metabolic studies are important in answering many questions, such as how a gene's mutation affects phenotypes of the plant, and what mutations cause diseases in a specific plant. Many biologists advocate metabolic profiling in the study of gene fuction (Dixon et al. 2006). In the 1990s, metabolic analyses in plants and animals were developed. In this project, we will study mutation effects in the plant *Arabidopsis thaliana*. (More detailed information on the 9 lipidomics experiments is given in Appendix A: The Lipidomics Experiment Information).

1.3.1. The Metabolome/Lipidome Data Analysis Work Flow

In a lipidomics experiment, biologists use several high-throughput technologies to identify and quantify the composition of samples from the organisms. Table 1.2 lists some of the MS data analysis techniques used in lipidomics experiments. In this dissertation our focus is on the targeted analysis experiment with some primary interest in particular lipid compounds that are to be analyzed.

Table 1.2: Some techniques used in metabolic studies (Dunn and Ellis, 2005; Hall, 2005; Fiehn, 2006; Nielse et al. 2005)

Terminology	Definition
Metabolite profiling	The methodologies to identify and quantify a group of specific metabolite compounds, e.g. such as carbohydrates and amino acids.
Metabolite target analysis	High-throughput technology to identify and quantify metabolites that participate in a specific part of metabolism.
Lipid profiling	A targeted metabolomics technique to identify and quantify lipid species with high sensitivity mass spectrometry, e.g. ESI-MS/MS.

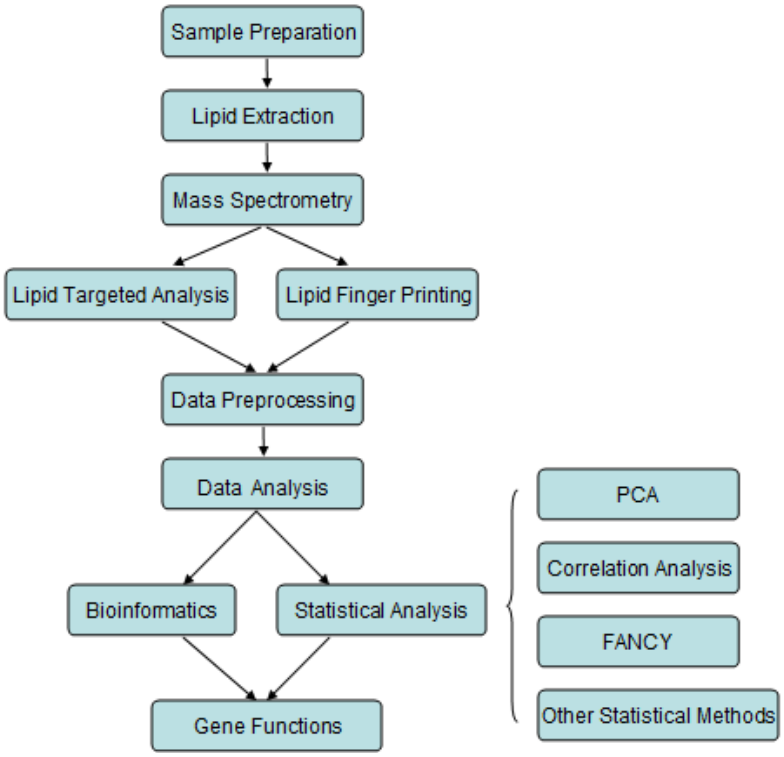

Figure 1.3: The lipid data analysis work flow chart (Welti and Wang, 2004)

Figure 1.3 shows the work flow of lipidomics data analysis. The electrospray ionization tandem mass spectrometry (ESI-MS/MS) was utilized in this experiment (see Table 1.1). The method used is a targeted analysis. There are several ways for researchers to find the functions of genes. One way is to start use a public databases (e.g., KEGG: *http://www.genome.jp/kegg/pathway.html*) to find a map the metabolic pathway. Lipid maps (*http://www.lipidmaps.org/*) provides another resource with rich information for understanding lipids, including techniques for exploring functions of lipids, lipid pathway mapping tools, and bioinformatics tools. Other methods include using statistical methods to find a list of lipid

combinations that are functioning together in a pathway. Among all the statistical methods, the most commonly used methods are PCA (Raamsdonk et al. 2001), correlation analysis (Fukushima et al. 2011), and functional analysis by co-responses in yeast (FANCY) (Raamsdonk et al. 2001). In chapter 2, the literature review, several techniques will be described for discovering lipids that are functioning in pathways or as reactant-product pairs.

1.3.2. The Lipidomics Pathways

The concentration of lipid metabolites may change due to both internal and external factors (Welti and Wang, 2004). While internal factors include the genetic activities, external factors include environmental conditions, such as freeze or drought stress. Therefore, the concentration level change in lipids not only can show the underlying biological reactions but also can reflect the enzymatic activities which can be used to make inference about the gene. The pathway can be described as a series of biological reactions that can form a long chain of reaction paths or reaction networks. Figure 1.4 is an example of the lipid pathway network.

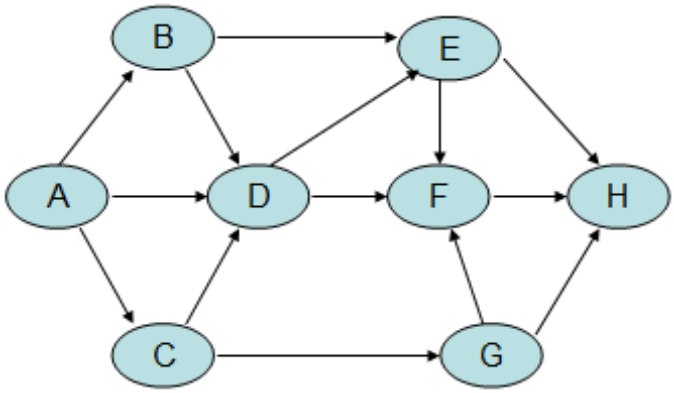

Figure 1.4: **A schematic drawing of the lipid pathway network**
Letters A, B, C, D, E, F, G and H are used to denote lipid compounds. The arrows are drawn from a reactant lipid compound to a product lipid compound.

In the pathway network described in Figure 1.4, the arrows indicate different reaction paths. For example, in the notation lipid A → lipid B, the reaction path is from lipid A to generate lipid B. The lipid A serves as the reactant (substrate) of the enzyme, and lipid B serves as the product of the enzyme. This notation, A as the reactant and B as the product, will be used throughout this dissertation. In Figure 1.4, the pathway consists of a series of reactions which include some lipids that can be both reactant in one pair and product in another pair. In this

dissertation we will explore any chain of reactions or reaction networks, by focusing on pairs A → B within the chains. The context will be only briefly discussed in some of the results.

1.3.3. The Fundamental Scheme Used in the Pathway Data Analysis

As mentioned above, we use the notation A to represent the reactant and B as the product for an arbitrary lipid reactant-product pair. All symbolic notations used herein are listed below, where WT = wild type and MT = mutant:

A	Reactant in the pathway
B	Product in the pathway
A_w	Reactant concentration level in WT
A_m	Reactant concentration level in MT
B_w	Product concentration level in WT
B_m	Product concentration level in MT
$\overline{A}_w$	Reactant mean concentration level in WT
$\overline{A}_m$	Reactant mean concentration level in MT
$\overline{B}_w$	Product mean concentration level in WT
$\overline{B}_m$	Product mean concentration level in MT

Figure 1.5 illustrates the scheme used to find a reactant and product lipid pair in the lipidomics pathway. (a) A → B is a general notation for a reactant and product pair if A is a reactant and B is its product in the pathway. (b) $A_w \rightarrow B_w$ is a notation to show that A_w can generate B_w. Since there is no mutation effect in the wild type organism, this reaction leads to a decreased concentration level of A_w and increased concentration level of B_w. Still in step (b), $A_m \not\rightarrow B_m$ is a notation to show that the generation of B_m from A_m is blocked if there is a mutation effect on the mutant organism that is affecting the pathway between reactant and product. This blockade happens because the mutation alters the level of the enzyme that is used to catalyze the reaction. As a result, the concentration level of the reactant A_m increases, but the level of B_m decreases. In general, if A is the reactant and B is the product in the pathway, both paths $A_w \rightarrow B_w$ and $A_m \not\rightarrow B_m$ in Figure 1.5(b) should be valid at the same time. Therefore, the reactant A should have higher concentration level in the MT than that in the WT, and the product B should

have lower concentration level in MT group than that in the WT group, leading to the two relations shown in (c), i.e., $A_w < A_m$ and $B_w > B_m$. A typical scatter plot of the lipid reactant and product pair AB is illustrated in Figure 1.6.

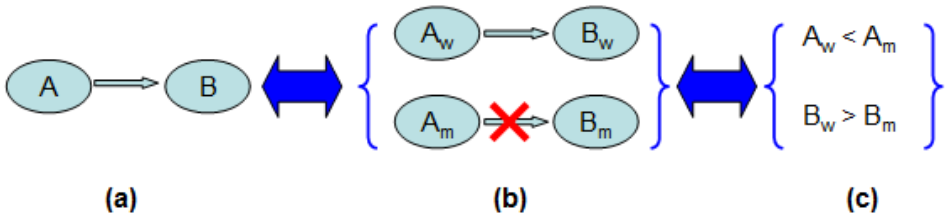

Figure 1.5: The scheme used to find reactant and product AB lipid pairs
If A is the reactant and B is the product in the pathway in **(a)**, then the concentration level of the reactant A can generate B in the WT, i.e., $A_w \rightarrow B_w$, but A cannot generate B in the MT group, i.e., $A_m \not\rightarrow B_m$ in **(b)**. As a result, the concentrations of A and B in the WT and MT leads to $A_w < A_m$ and $B_w > B_m$ in **(c)**, (Fan 2010).

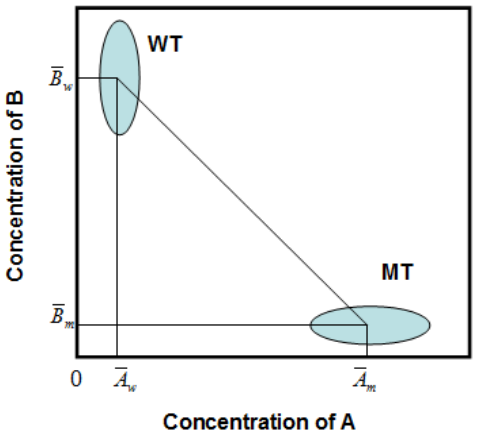

Figure 1.6: The concentration level relationships of reactant and product AB lipid pairs
This figure illustrates our scheme, $A_w < A_m$ and $B_w > B_m$. In the data from this experiment, WT group includes 5 data points of (A_w, B_w). The MT group includes 5 data points (A_m, B_m), where A and B stand for the concentration levels in the reactant A lipid and product B lipid, respectively.

Since the mean is a common summary statistic usually used in statistical analysis, we use sample means, i.e., $\bar{A}_w < \bar{A}_m$ and $\bar{B}_w > \bar{B}_m$ which are also shown in Figure 1.6 to represent the scheme.

1.4. Dissertation Outline

Chapter 2 reviews the literature on metabolite pathway analysis, with several main methods used in recent years, namely, PCA, FANCY (Functional Analysis by Co-response in Yeast by Raamsdonk et al. 2001) and correlation analysis. The results from those methods are compared with our screening scheme.

Chapter 3 explores characteristics of the lipid experiment targeted analysis datasets. All the physical characteristics from a selected set of biologically functional lipid pairs (to be described later) will be summarized and used to discriminate the arbitrary lipid pairs. Summary test statistics are derived from the lipid pairs that are of interest as a result of exploratory data analysis. We find that the test statistics from these lipid pairs can capture all the structure in the scheme shown in Figure 1.5 and Figure 1.6.

Chapter 4 outlines a method to fit a null distribution using the bootstrap. The bootstrap null distributions of each statistic are derived using the Parametric Bootstrap Null (PBN) distribution with Kolmogorov-Smirnov (K-S) test statistic.

Chapter 5 describes an alternative way to fit a bootstrap null distribution obtained in chapter 4 using a mixture of normal distributions. The bootstrap null distributions are found by using a two-component mixture model which can capture the shape of the empirical bootstrap null distribution.

Chapter 6 investigates the performance of the two different methods, MNBN and PBN, under a different null hypothesis, that is, one of equal means rather than equal distribution as was done in chapter 4 and 5. Two datasets are used as examples to compare the results obtained under the null assumptions in chapter 4 and chapter 5, respectively.

Chapter 7 explores another mixture model approach by using a normal mixture model to fit a statistic from chapter 3, called R_T. The number of components in the mixture model is tested. Since larger R_T values indicate significant results, the second normal component in a two-component mixture model or the third normal component in a three-component mixture model, which is used to model larger values of R_T, will produce the posterior probabilities for a lipid pair that is affected by the mutation at a given R_T value.

Chapter 8 explores the relation between the concentration levels of A and B lipids and the role of a lipid being a reactant A or product B in a lipid pair by using a two-way ANOVA model with an interaction term. Since the interaction term in a two-way model can explain the screening

scheme relationships, the fixed interaction effect will be tested for each lipid pair to check if the pair is significantly affected by the mutation or not.

Chapter 9 simulates some realistic data that can be used to evaluate the performance of the new statistical methods in the area of lipidomics. One of the methods, Mixture Normal Bootstrap Null (MNBN) distribution, is used as an example and applied to a simulated dataset which is generated based on a real dataset.

Finally, a summary of the dissertation and some new directions for future work will be illustrated in chapter 10.

Chapter 2 - Literature Review

2.1. Pathway Analysis Overview

In the 1990s, metabolic analysis in plants and animals was developed by using comprehensive, sensitive high-throughput mass spectrometry (MS). Many techniques for analyzing high-dimensional metabolome data have been invented since then. Plant biologists started to use MS analysis in the area of lipidomics profiling in 2002 (Welti et al. 2007; *http://en.wikipedia.org/wiki/Lipidomics*). Research interests in lipidomics profiling focus on the roles of lipids in the cell, and lipid responses to stress conditions, such as temperature, salinity, sulfur, phosphorus, heavy metal, and oxidative conditions (Welti et al. 2007; Shulaev et al. 2008). Among all interests, a very interesting research area is lipid pathway/network analysis which involves identifying the functions of a mutated gene by comparing metabolite data of wild-type and mutant organisms.

Lipidomics research has advanced rapidly in the past ten years. Various tools and databases have been built which have become useful resources in a metabolite or lipid pathway, such as MetaCyc (Caspi et al. 2006), MetNetMaker (Forth et al. 2010), PathExpress (Goffard et al. 2009) and MedicCyc (Urbanczyk-Wochniak et al. 2007). Lipid Maps (*www.lipidmaps.org*) is a lipid data repository for investigating lipid structure and classification. Another public data base, KEGG (*www.genome.jp/kegg/*), is not only used as a repository for gene data, but also provides more information for analysis of metabolic pathways and of the interaction between genes and metabolites. Online resources, such as Cyberlipid (*http://www.cyberlipid.org/*) and Lipid Bank *(http://lipidbank.jp/*) provide valuable references for analyzing and interpreting lipid data.

In pathway analysis, many approaches can be used to achieve the same goal in finding metabolite networks. The main principle used in all methods in the literature for detecting metabolites in the same pathway is to measure and analyze changes in concentration. The change in concentration of metabolites is the result of an underlying biochemical process. Different techniques emphasize different aspects in finding a list of significant metabolite pairs or to find a pathway network. As shown in Figure 1.5, our experimental scheme can be summarized into two expressions: $A_w < A_m$, and $B_w > B_m$. This scheme not only can emphasize the lipid level change in the wild type and in the mutant groups, but also can characterize specifically which lipid is the

reactant and which lipid is the product. Determining the reaction direction from the reactant A to the product B, A→B, is the main point in this study. Identification of reactant-product pairs has not been considered in most other techniques to be discussed in this chapter. Figure 2.1 compares the main ideas of some primary methods used in this research area.

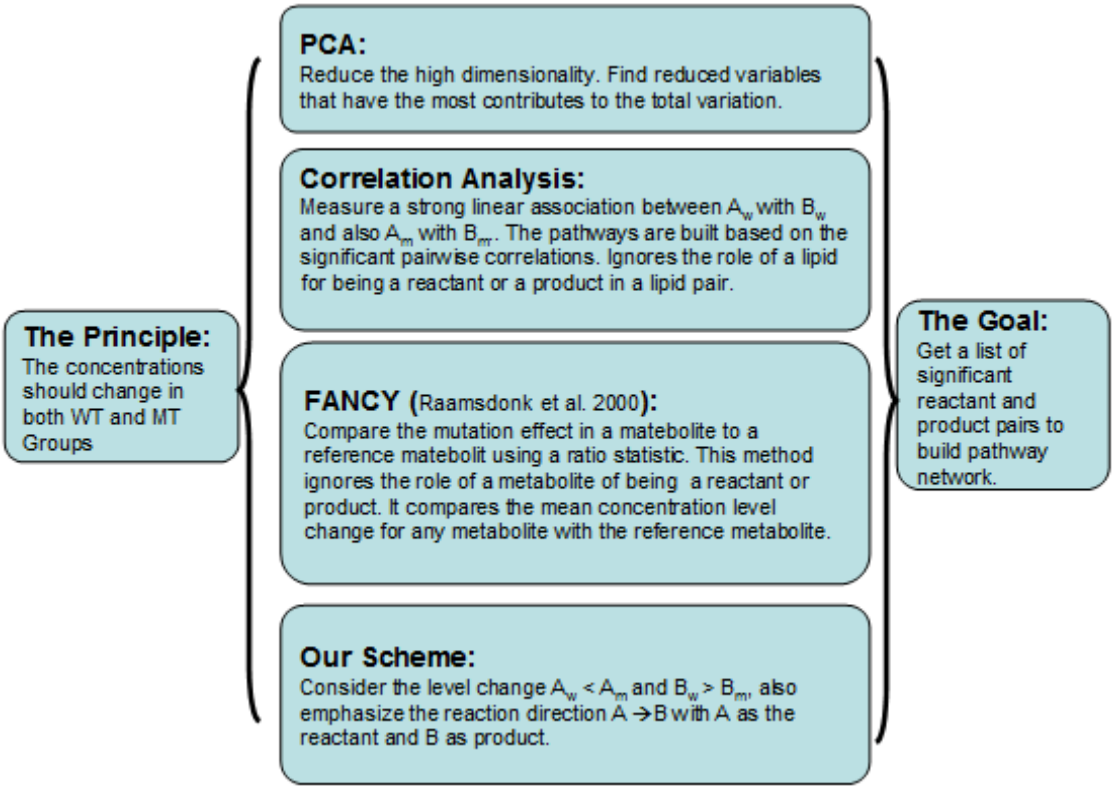

Figure 2.1: Comparison of some main methods for detecting a metabolite pathway

In Figure 2.1 Principal Component Analysis (PCA) (Raamsdonk et al. 2001; Fukushima et al. 2011) is an unsupervised approach that is widely used in high-dimensional biological data analysis, such as in microarray data, protein data and lipid data. Other unsupervised learning techniques, e.g, hierarchical clustering Analysis (HCA), and supervised learning techniques, e.g, partial least squares (PLS), are also commonly used by researchers. The PCA techniques can reduce the high dimensionality (e.g., variables, or lipid species). It is usually used as a pre-analysis step for other discriminant analysis techniques.

PCA is also a useful visualization tool (*http://ordination.okstate.edu/PCA.htm*) for multivariate analysis when the genes or lipid species are used as variables. PCA is used to identify underlying factors which explain the correlations among a set of variables. The orthogonal

principal components can be reduced to several factors (or principal components), and then the factors are used to explain correlations and differences in the lipid species. Since PCA is not a method to find the metabolite pathway directly, this will not be investigated here.

As shown in Figure 2.1, correlation analysis (Weckwerth et al. 2004; Fukushima et al. 2011; Görke et al. 2010; Steuer et al. 2003; Steuer 2006; Allen et al. 2010) is used in pathway analysis by using an assumed linear association between a metabolite X with another metabolite Y in WT and MT groups separately, i.e. between X_w vs Y_w or X_m vs Y_m. Sakurai et al. (2011) built a gene-to-gene and metabolite-to-metabolite pathway database, KaPPA-View4 (*http://kpv.kazusa.or.jp/kpv4/*), using a correlation network. Morgenthal et al. (2006) introduced correlation (heat) maps to show the network structure by using pair wise correlations from different parts (leaf and tuber) of the plant, *Arabidopsis thaliana.*

Since metabolome or lipidome pathway analysis involves many biological algorithms, some biologists found metabolite networks by using their expertise in biology, but with less application of statistical analysis. Wu et al. (2005) introduced their approach using a metabolic flux (e.g., the growth rate) model. The FANCY approach (Functional Analysis by Co-responses in Yeast) performed by Raamsdonk et al. (2001) used single-celled yeast to find co-responses of metabolites that were caused by mutation effects (see Figure 2.1). FANCY results are derived and based on a metric that is an angle and results are interpreted based on the value of this angle, rather than doing a statistical test to assess significance. Since the FANCY method is one of a few related to the work described herein, more details of this approach will be discussed in a separate subsection.

In high-dimensional data analysis, multiple testing procedures are used to control the Family-Wise Error Rate (FWER, the chance of making one or more Type I errors among multiple tests of hypotheses), such as Tukey-Kramer's HSD (Honestly significant difference) test for pair-wise comparisons, Dunnett's test for pair-wise comparisons between the control group versus all other treatment groups, and Scheffé's method for testing all possible contrasts between treatment means. The Bonferroni correction divides the statistical significance level by the total number of lipid features and this can be too stringent for controlling the type I error rate when there are a large number of lipid variables in a lipid dataset. False Discovery Rate (FDR) control is commonly used in "omics" high-dimensional data analysis with a large number of hypotheses tests. Some brief remarks about FDR related methods will be given in a later subsection.

2.2. FANCY Approach

Introduction to FANCY Method

Raamsdonk et al. (2001) introduced a technique to find the function of "silent" genes using metabolite level changes from a single-celled organism, *Saccharomyces cerevisiae*, which is a species of yeast. The "silent" genes are those genes with functions that are not reflected in the visible or obvious phenotype. By deleting "silent" genes from the *Saccharomyces cerevisiae*, the researchers expect to reveal the role of unknown genes by comparing them to the genes of known function using a co-response coefficient in the FANCY approach (Functional Analysis by Co-responses in Yeast).

The FANCY approach was developed based on changes in metabolite concentration levels. Raamsdonk et al. (2001) hypothesized that it is possible to identify gene sites by looking at the metabolite concentration level change that is caused by the mutation. The authors demonstrated their FANCY technique by using one WT treatment and 4 mutant treatments and they focused on just a total of 6 metabolites species plus one reference metabolite, G6P(i.e., glucose-6-phosphate). WT and mutant type are actually different genotypes of an organism. For convenience and to highlight that it is the effect of the mutation on the lipidome that is of interest, herein they are referred to as 'treatments' even though it is not a treatment in the sense that it is something that can, technically, be randomly applied to an experimental unit. The co-response from Raamsdonk et al. is defined by Ω:

$$\Omega_{mutation}^{X:G6p} = \frac{\ln(X_m) - \ln(X_w)}{\ln(G6P_m) - \ln(G6P_w)}. \qquad (2.1)$$

X refers to the mean concentration level of any of the 6 metabolites with 3 samples in the WT and MT groups. The subscripts *m* and *w* denote the wild type and mutant type treatments, respectively. For this review of literature some notational convenience will be employed that has, in fact, been used in publications such as the one being summarized here that presented equation (2.1). Variables X and Y will refer to concentration levels of two different metabolites (or lipids). Sometimes we will just state metabolite X and Y to refer to concentration levels of two metabolites or mean concentrations. Thus "X" will be used both to label a metabolite and to denote its concentration level. Use of a subscript, such as X_w, will denote the same but for a WT organism. It is hoped that the meaning of the terminology will be clear from the context in which it is used.

The numerator of the co-response coefficient shows the log mean concentration level change between the MT and WT for a metabolite X. The denominator shows the log mean concentration level change between the MT and WT groups for the reference metabolite G6P. Since the co-response Ω is a ratio statistic, it has the disadvantage of being infinity if the change in the denominator is very small. Raamsdonk et al. (2001) converted the co-response into an angle response $\theta_{mutation}^{X:G6p}$ by letting

$$\theta_{mutation}^{X:G6p} = \arctan(\Omega_{mutation}^{X:G6p}) \tag{2.2}$$

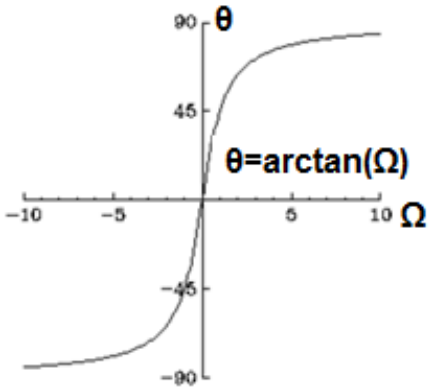

Figure 2.2: Plot of the co-response θ as a function of Ω.
Image from: *http://thesaurus.maths.org*

The co-response coefficient $\theta_{mutation}^{X:G6p}$ in (2.2) is an angle measurement. The range of θ is $-90° < \theta < +90°$. If $\theta > 45^{\circ}$ or $\theta < -45^{\circ}$, from (2.1) we can see the mutation effect is larger in the X metabolite than that in the reference metabolite G6P. On the other hand, if $-45^{\circ} < \theta < 45^{\circ}$, the mutation effect is smaller in the X metabolite than the reference metabolite G6P.

The FANCY approach can be summarized into the following steps.

1) Pair metabolites using all possible metabolites, X, with the reference metabolite G6P to get 6 pairs (recall, the Raamsdonk et al. focused on only six selected metabolites). Since there are in total 6 metabolites and one reference metabolite G6P, 6 pairs of metabolite were analyzed. The 6 metabolites are denoted $X_1, \ldots, X_6$ in what follows.
2) For each metabolite pair, get the co-response coefficient Ω in (2.1).
3) Convert the co-response coefficient Ω into its angle response θ in (2.2).

Table 2.1 illustrates the FANCY results. In the table, the angle responses are listed for all metabolite pairs under 4 different mutant treatments. For example, the angle response θ with

mutant 1 applied on metabolite pair X_1/G6P is +80°, larger than +45°, then the co-response $\Omega > +1$. From equation 2.1, this means mutation 1 resulted in a larger change in the concentration of X_1 versus the WT than it did the reference metabolite across mutant and WT. Therefore, mutation 1 had a larger effect on the metabolite X_1 than that in the reference metabolite G6P. The positive signs of θ and Ω mean that the concentration levels change in the same directions in both the metabolite X_1 and the reference metabolite. Similarly, the angle response θ with mutant 4 applied on metabolite pair X_1/G6P is $-60°$, and the co-response $\Omega < -1$. From equation 2.1, this means mutation 4 causes a larger change in the concentration of X_1 than in the reference metabolite. Therefore, mutation 4 has larger effect on the metabolite X_1 than in the reference metabolite G6P. However, here the negative signs of θ and Ω mean that the changes in the concentration levels of the metabolite X_1 and reference metabolite G6P are in opposite directions.

Table 2.1: Example results of the FANCY method shown with the angle variable θ
In the first row, X_i/G6P stands for the pairs formed by any one of metabolites and the reference metabolite, G6P, where i = 1, …, 6. The four mutant treatments are listed in the first column. The general notation "Mutant j" are used instead of the real mutant name to avoid complicated biological terms, where j = 1, 2, 3, 4. The column names and row names are not the real names from Raamsdonk et al. The symbols are used as displaying purpose. The angle values are from Raamsdonk et al. (2001) Table 3, p 49.

MT treatments	X_1/G6P	X_2/G6P	X_3/G6P	X_4/G6P	X_5/G6P	X_6/G6P
Mutant 1	+80°	+80°	−60°	+80°	−60°	−80°
Mutant 2	+60°	+50°	−30°	+60°	−50°	−70°
Mutant 3	−70°	−80°	+80°	−80°	−60°	+80°
Mutant 4	−60°	−80°	+80°	−80°	−70°	+90°

Comparison between FANCY and Our Screening Scheme

Our screening scheme can be summarized into two relationships $\overline{A}_w < \overline{A}_m$ and $\overline{B}_w > \overline{B}_m$ as described in chapter 1. We pair all arbitrary lipid pairs and discard all other lipid pairs that do not satisfy the screening scheme. In our analysis, we specifically identify the reactant A and product B in each reaction in the pathway.

In the FANCY approach, metabolite pairs are not paired arbitrarily. All metabolite Xs are compared with a reference metabolite G6P. In the metabolite pair, the reaction direction is not clear and the roles of reactant and product are not specified. The FANCY method was applied to

a very small dataset with 6 metabolite species using 3 samples. The angle values are an indication of the mutation effects in each metabolite pair. The FANCY method ranked lipid pairs by the angle metric and did not offer any test of significance.

2.3. Correlation Analysis

2.3.1. Introduction to Correlation Analysis

Weckwerth et al. (2004) identified metabolic networks by using correlation analysis. They assumed that the metabolic fluctuation in the WT and MT groups might have a linear association between metabolite concentration levels. They also emphasized that the concentration level changes in metabolites was evidence of an underlying pathway structure and built a pathway network by using the following steps.

1. They found significant mean differences for metabolite concentrations between the wild type and the mutant type groups by using a t-test and adjusting the FWER by using a Bonferroni adjustment.
2. Genotype discrimination was performed by PCA and DFA (Discriminant Function Analysis) to see what metabolites species show differences between groups.
3. Let X and Y be an arbitrary metabolite pair without specifying the reactant and product. Pearson's correlation analysis was applied to each metabolite pair to see if there was a significant linear association between X_w and Y_w (or between X_m and Y_m). The threshold for significant correlations is 0.8. Slopes of the lines connecting X_w and Y_w, or X_m and Y_m from a regression model were also used as a statistic to identify metabolite pairs that are significantly affected by the mutation.
4. Pathway networks were constructed for WT and MT groups separately based on the list of significant results from correlation analysis.

Fukushima et al. (2011) used a slightly different correlation analysis approach to find metabolite networks by using correlations of sample data between a pair of metabolites X_w and Y_w in the WT, $r_{x_w y_w}$, and X_m and Y_m in the MT group, $r_{x_m y_m}$. Fukushima also analyzed the correlations within the same treatment from different parts of the plant, i.e., using X_w with $X_{w'}$ or X_m with $X_{m'}$ ($r_{x_w x_{w'}}$ or $r_{x_m x_{m'}}$) for the same metabolite species.

2.3.2. Roots-aerial Datasets Correlation Analysis in Fukushima et al. (2011)

Fukushima et al. (2011) performed the correlation-network for metabolites by using roots and aerial tissues of the modeling plant *Arabidopsis thaliana* by using high-throughput chromatography-time-of-flight/mass spectrometry (GC-TOF/MS). The authors applied correlation analysis to two datasets and used a clustering algorithm to build a metabolite network. One dataset was from the roots data from Fukushima et al. (2011). Another dataset was the aerial data from Kusano et al. (2007).

In the two datasets, there were three treatment groups with *Col-0* as the wild type (WT) and *methionine over-accumulation* 1 (*mto1*), and *transparent testa4* (*tt4*) as the two mutant groups. 59 metabolites (variables) which commonly exist in both roots and aerial tissues were analyzed. The sample sizes used in those two datasets are listed in the following table.

Table 2.2: The sample sizes in roots-aerial datasets

	WT	**mtol**	**tt4**
Roots	17	16	20
Aerial	17	13	20

Fukushima and colleagues analyzed the roots-aerial data in the following procedure.

1) The Spearman correlation was used for the correlation analysis. Sample pair-wise correlations of X_w and Y_w from the WT ($r_{x_w y_w}$), and X_m and Y_m from the MT group ($r_{x_m y_m}$) were tested as follows: let ρ be the population correlation between metabolite X and Y in a pair. Using WT as an example, test $H_0 : \rho_{x_w y_w} = 0$, $H_a : \rho_{x_w y_w} \neq 0$ using

$$t_{stat} = r \cdot \sqrt{\frac{n-2}{1-r^2}} \xrightarrow{t} t_{(df=n-2)} \tag{2.3}$$

where *r* denotes the corresponding Spearman's correlation coefficient calculated from sample data obtained for the two metabolites.

2) For any two correlations in Table 2.3, the correlation difference between any two metabolites, denoted $\rho_1 - \rho_2$, was tested using the corresponding sample correlation difference for those two metabolites, denoted $r_1 - r_2$. Fisher's Z transformation was applied to the correlation difference $r_1 - r_2$ with the hypotheses $H_0 : \rho_1 = \rho_2$ vs. $H_a : \rho_1 \neq \rho_2$ based on the test statistic

$$Z_{stat} = \frac{0.5\ln\left(\frac{1+r_1}{1-r_1}\right) - 0.5\ln\left(\frac{1+r_2}{1-r_2}\right)}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}} \quad (2.4)$$

where Z_{stat} is from a Z test of two sample correlation difference in the Fisher's transformation.

Table 2.3: The sample correlations in roots-aerial datasets
Each table cell stands for a vector of correlations between a metabolite pair X and Y in the same treatment. Each dataset has the same number of 1711 common pairs.

	WT	**mtol**	**tt4**
Roots	$r_{x_w y_w}$	$r_{x_{mtol} y_{mtol}}$	$r_{x_{tt4} y_{tt4}}$
Aerial	$r_{x_w y_w}$	$r_{x_{mtol} y_{mtol}}$	$r_{x_{tt4} y_{tt4}}$

The following correlation analyses were applied to the roots-aerial datasets by using the above algorithms.

1) Applied Log-transformation on all 6 datasets in Table 2.2 with 59 common metabolites in each dataset.
2) Paired all the arbitrary metabolite X and Y in each dataset to make $\binom{59}{2} = 1711$ common pairs.
3) Calculated the Spearman's correlations for each pair in each dataset to get 6 sets of sample correlations in Table 2.3.
4) Assessed the significance of correlations for each correlation dataset using t test in equation (2.3). Used local *fdr* to adjust for multiple testing. Compared the mutation effect by using the number of significant correlations in each of the treatments WT, *tt4* and *mtol*. The comparison plots are shown in Figure 2.3 A.
5) All correlation differences from Table 2.3 were tested using the normal approximation test statistic in (2.4). They compared, in particular, the correlation differences within each genotype (WT, *tt4* and *mtol*) between roots and aerial tissues, and then used the local FDR for multiple adjusting testing adjustment a threshold of 0.05. The differential correlation comparison plot is shown in Figure 2.3 B.

2.3.3. Results for Roots-Aerial Datasets

Fukushima et al. (2011) examined the number of significant correlations between tissues in each treatment from both roots and aerial datasets. Figure 2.3A shows a comparison of the significant correlations in each genotype/treatment for the roots and aerial tissues. Figure 2.3B shows the significant differential correlations by using $r_1 - r_2$ for each genotype between roots and aerial tissues. From Figure 2.3A, we can see that the number of significant correlations in WT and mtol is bigger in root tissues than in aerial tissues. But tt4 has a bigger mutation effect in aerial tissue than in roots. Figure 2.3B shows that tt4 has more differential correlations than the other two genotypes. Table 2.4 shows the number of significant pairs in the three genotypes/treatments which are visualized in Figure 2.3 A. Table 2.4 lists the results for Figure 2.3.

Colleagues of Fukushima, Kusano et al. (2007) performed correlation analysis using the same aerial dataset with more metabolite species. Kusano et al. found the common metabolite correlations in all the three genotypes using a Venn diagram.

Table 2.4: Number of significant correlations in each tissue and genotype/treatments
This table is produced by following the description from Fukushima et al. (2011). For example in the WT group in the roots, the number of significant correlations between $X_{w,roots}$ vs $Y_{w,roots}$ is 250. 170 significant correlations between $X_{mtol,roots}$ vs $Y_{mtol,roots}$ were found in the roots in mutant *mtol*. 413 significant correlations between $X_{tt4,roots}$ vs $Y_{tt4,roots}$ were found in the roots in mutant *tt4*.

	WT	mtol	tt4	Total
Roots	250	170	413	833
Aerial	150	104	565	819

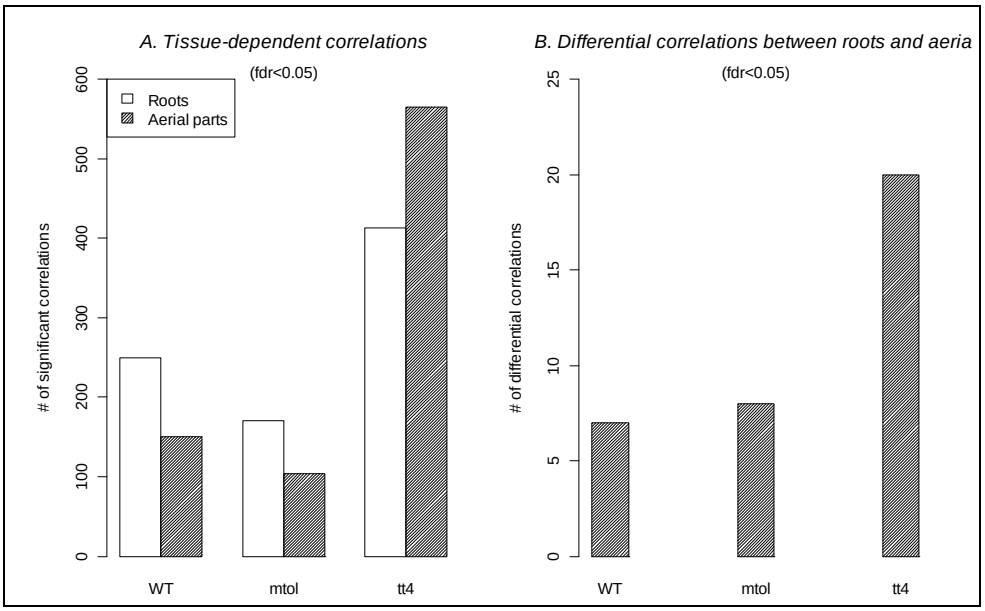

Figure 2.3: Correlation analysis results for comparing the 3 genotypes in *Arabidopsis*
(A) The number of significant correlations in each of the three treatments in roots and aerial tissues is shown in Table 2.4. For example in the WT group, the significant correlations between $X_{w,roots}$ vs $Y_{w,roots}$ and $X_{w,aerials}$ vs $Y_{w,aerial}$ is compared. X and Y stand for one metabolite pair. (B) The number of significant correlations comparing the same genotype/ treatment from roots and aerial tissues, e.g., in the WT treatment, the number of significant correlation difference $r_1 - r_2$ is shown, r_1 is a correlation between $X_{w,roots}$ and $Y_{w,roots}$ and r_2 is a correlation between $X_{w,aerials}$ and $Y_{w,aerial}$. X and Y are the same pair in r_1 and r_2. (This figure is re-produced from the data used in Fukushima et al. (2011) and corresponds to Figure 3 in their paper.

Figure 2.4 shows the numbers of common correlations from three genotypes which is re-produced from the roots-aerial data from Fukushima et al. (2011) using the Venn diagram to show the relationships of the significant metabolite pairs with the three genotypes. Figure 2.4 (a) is a visualization of a total 833 (in the first row of Table 2.4) significant correlations in the roots data from three genotype: WT, *mtol* and *tt4* of which 87 significant correlations are from the WT group alone, 222 correlations are from treatment *tt4* alone and 36 significant correlations are from only treatment *mtol*. In addition, 62 significant correlations commonly exist in all the three treatments. 79, 22, and 50 correlations exist in pair wise treatment intersections. The intersections of the metabolite pairs appeared in more than one treatment. Also note that in the first row of Table 2.4 in the roots in the WT, 250 significant correlations are from the sum of 87, 79, 62 and 22. The total 833 significant correlations include the duplicated pairs if the same pair appeared more than one time on the list. Figure 2.4 provided more information than Figure 2.3

on what metabolites species are significantly affected by more than one treatment if they are in the intersections of the list.

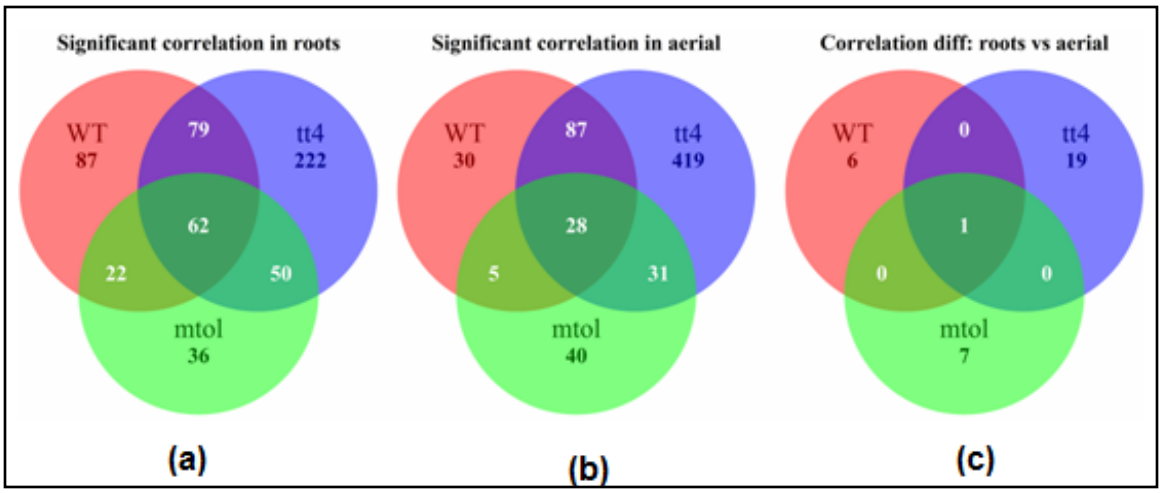

Figure 2.4: The numbers of common correlations from the three genotypes
The three diagrams are re-produced from roots-aerial datasets from Fukushima et al. (2011), and follow the description from Kusano et al. (2007). (a) and (b) are Venn diagrams of the three genotypes in the roots tissues with the significant correlations converted from Figure 2.3(A). (c) Venn diagram of significant correlation differences between the roots and the aerial in the same genotypes which are converted from figure 2.3(B).

Compared with our screening scheme showed in Figure 1.5, Fukushima et al. focused on the linear association between a metabolite pair X and Y in WT or in MT groups separately, and were interested in correlation changes among metabolite pairs across roots and aerial tissues within the same genotype. The correlation analysis focused on the strong linear association in each treatment groups which did not discriminate the reactant and the product role for each metabolite pair like what we do in our screening scheme.

Weckwerth et al. (2004) also focused on the correlation differences from different treatment groups. They demonstrated their correlation analysis using slopes in the regression analysis for finding the relation between the Y metabolite and the X metabolite in a metabolite pair. The main idea of the analysis is illustrated in Figure 2.5 by using the roots-aerial dataset from Fukushima et al. (2011). In a centered and scaled dataset, the slope is the same as the correlation in the same pair which will be discussed in chapter 3.

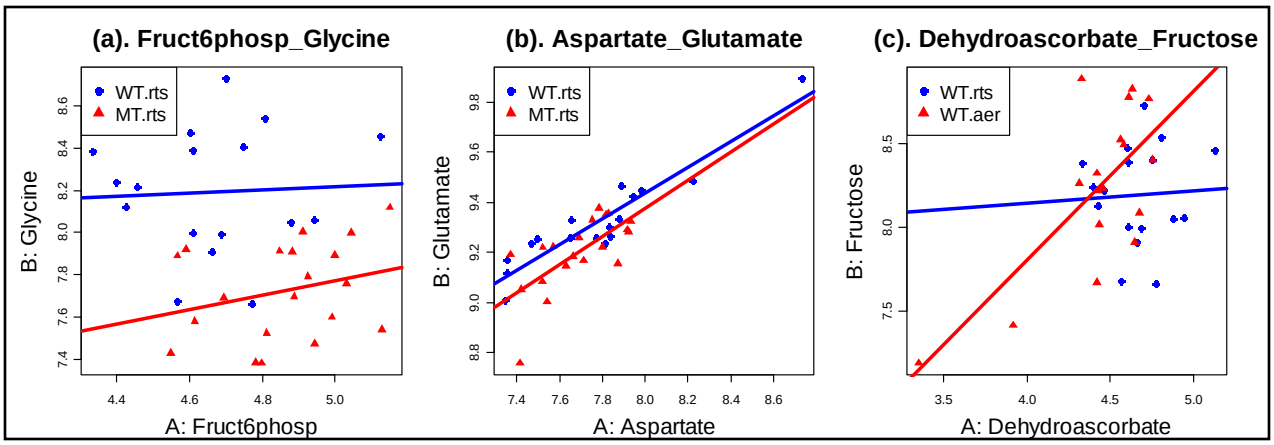

Figure 2.5: Illustration of the correlation analysis using slopes in each treatment group
(a) Blue circles represent the WT in roots. Red triangles are the mutation *tt4* in roots. The blue line is the regression line from X_w with Y_w in WT groups. X and Y stand for a metabolite pair. The red line is the regression line for fitting X_m with Y_m in the mutant *tt4* treatment. (b) The same symbols with (a) are used, but the two slopes are significant slopes in both WT (blue line) and significant in the MT (*tt4* with red line), but not significant in the correlation difference (between the two line). (c) The blue circles are the WT in roots data and the red triangles are the WT in aerial data for the pair Dehydroascorbate and. Fructose. This plot is for a significant correlation difference between the red line and the blue line.

Figure 2.5(a) shows a typical scatter plot with the regression line in each group by using a metabolite pair Fructose-6-phosphate and Glycine. The red and blue regression lines are fitted by using concentration levels of X and Y in two treatments WT and MT, *tt4*, in the roots data, i.e., the red line is from X_{tt4} with Y_{tt4} and the blue line is from X_{wt} with $Y_{wt.}$, where X and Y metabolites refer to Fructose-6-phosphate and Glycine, respectively. The two slopes are not significantly different from 0.

Figure 2.5(b) displays the slopes or the correlation relations from a metabolite pair that shows a significant WT correlation from Fukushima's correlation analysis results. A metabolite pair between Aspartate and Glutamate has the most significant correlation in the WT dataset (with the blue regression line) with the *lfdr* = 9.82×10^{-6}. The *tt4* group correlation is also significant with *lfdr* = 0.01599 (for the red line). But the difference between the two slopes (correlations) are not significant with *lfdr* = 0.551013.

Figure 2.5(c) shows a significant correlation difference $r_{wt,aerial} - r_{wt,root}$ in the WT between aerial and roots correlations for the metabolite pair Dehydroscorbate and fructose. The *lfdr* (local *fdr* defined in the next section) for the correlation difference is 0.004994. The

correlation for the WT in the aerial data is significant with *lfdr* = 0.003702. But the correlation in the WT in roots data with *lfdr* = 0.23863 is not significant.

In the pathway analysis, since a correlation difference between the WT and MT groups can reflect the mutation effect in each metabolite pair, the correlation difference should be a more important and interesting measure than if we only consider the significant correlations within each treatment group alone. Even though the correlation difference can show some interesting results, the significant linear association relationship may not include all other forms of relations. Non-linear relationships or other more complex relationships may need to be considered.

2.4. False Discovery Rates and Mixture Model Approaches in High-dimensional Data Analysis

The false discovery rate (FDR) is a commonly used method for controlling the number of type I errors in high-dimensional data analysis. Mixture model approaches have also been a useful statistical technique to find a list of significant genes that are affected by treatment conditions. A mixture model approach will be explored in this study to identify lipidomic pathways.

One could argue that elements of FDR theory go back to Schweder and Spjøtvoll (1982). However, it was developed into a more rigorous theory by Benjamini and Hochberg (1995). Table 2.5 illustrates the terms used in the FDR setting. “A” is the number of true negatives and “B” is the number of false negatives, the latter being the number of type II errors. “D” is the number of true positives and “C” is the number of false positives, the latter being the number of type I errors. The false discovery rate is defined as $\text{FDR} = E\left(\frac{C}{R}\right)$ when $R > 0$ and 0 if $R = 0$. FDR can be interpreted as the expected proportion of false positives among all significant hypotheses, and it is used to set an adjusted threshold for significance when thousands of hypotheses are tested simultaneously.

Table 2.5: Hypotheses testing terms for FDR in the high-dimensional data analysis.
K is the total number of hypotheses tested, and K = A + B + C + D. R is the number of the hypotheses rejected, and R = C + D.

	H_0 is true	H_0 is false	Total
Declare non- significant	A	B	K - R
Declare significant	C	D	R
Total	M_0	M_1	K

The local false discovery rate was introduced by Efron et al. (2001). Efron defined the local FDR, *lfdr*, at a particular value of a test statistic such as a z score as $lfdr(z) = \frac{p_0 f_0(z)}{f(z)}$, where

$$f(z) = p_0 f_0(z) + p_1 f_1(z), \tag{2.5}$$

is the density function for the data. The data in $f(z)$ consist of two portions. One portion represents the null component distribution, $p_0 f_0(z)$, and the other portion, $p_1 f_1(z)$, represents a distribution for which the alternative hypothesis is true. The proportion p_0 represents a proportion of tests for which the null is true while p_1 is represents the proportion of tests for which the alternative is true. The local FDR, $lfdr(z)$, is interpreted as a posteriori probability that a test is a true null given that it was rejected with a test statistic equal to z. More reviews on FDR methodologies can be found in Gadbury et al. (2008) and Broberg (2005).

Allison et al. (2002) used uniform null distribution for fitting a mixture model to the p values. Efron (2004) introduced the empirical null distribution frame work in a large-scale hypotheses testing. In simultaneous hypothesis testing with N null hypotheses, Efron converted the set of p values into their corresponding z scores by letting

$$z_i = \Phi^{-1}(P_i),\ \mathrm{i} = 1, 2, ..., \mathrm{N},$$

where z_i is the percentile for the i^{th} p value from the i^{th} test for that gene, Φ stands for the standard normal cumulative distribution function (cdf), Φ^{-1} represents the inverse standard normal cumulative distribution function, and P_i is the p value from the i^{th} test. Since z values are converted from a normal distribution, Efron named the *theoretical null distribution* of z_i as $f_0(z_i \mid H_i) \sim \mathrm{N}(0,1)$. The set of z_i was plotted into a histogram and then was used to get a natural spline curve for $f(z)$ by Poisson regression. Efron estimated the center (or mean), δ_0, by

using the relation $\delta_0 = \arg\max\{f(z)\}$ and standard deviation σ_0 for the spline curve $f(z)$. The standard deviation σ_0 is defined as

$$\sigma_0 = \left[-\frac{d^2}{dz^2} \log f(z) \right]_{\delta_0}^{-\frac{1}{2}}. \tag{2.7}$$

Efron named the null distribution $f_0(z) \sim N(\delta_0, \sigma_0)$ the *empirical null distribution*.

Efron compared the theoretical null distribution, $\mathrm{N}(0,1)$, with the empirical null distribution, $N(\delta_0, \sigma_0)$, in a gene expression study. It was shown that the empirical null distribution has some advantages over the theoretical null distribution. There may have many other null distribution methods in the mixture model fitting literature for gene expression studies. However, methods for metabolite pathway analysis using model fitting are less established. Nevertheless, the need for an empirical null distribution of test statistics defined in chapter 3 is just as relevant. We will explore approaches to find a null distribution by using the bootstrap procedure as described in chapter 4, and a normal mixture model approach to fit the null distribution in chapter 5.

Chapter 3 - Exploratory Data Analysis

3.1. Difficulties in Analyzing the Lipid Pathway Experimental Data

A total of $2\binom{141}{2} = 19740$ lipid pairs from the 141 lipids are considered in 9 different lipidomic experiments. Our research goal is to find the significant lipid reactant and product pairs that are affected by the mutation in the pathway. Methods should be applicable for detection of reactant product pairs in a metabolome through use of different genotypes or possibly different treatment conditions. Here, we need to refine the supporting evidence from the mutation effect on the A→B reactant-product pathway into a statistic. We must take into account non-normality, potential nonlinear relations, and also zero values in the datasets. A condensed version of the results in this chapter will appear in Zheng et. al (2013).

Dealing with Zero Values: There are some unique characteristics in the lipid pathway data which make the data analysis process more complicated. The difficulties are due to some zero values in lipid concentrations. There are 5 samples in each WT and MT treatment group. Some lipids have all zero values across the two treatment groups. Some other lipids may have all zero values in one treatment but not in the other. There are also some lipids with near zero concentration levels. Those zero values do not necessarily mean that the lipid is not present in the sample. Instead, the concentration level may be too low to be detected by the mass spectrometer. In this dissertation, the lipid is deleted from the data if its concentration is zero in all samples across both treatment conditions (or its standard deviation across all samples in both treatment conditions is zero). If the lipid has all zeros in one treatment but not in the other, it is kept. This kind of lipid could be a good candidate for a reactant or product.

Small Sample Size: Raamsdonk et al. (2001) analyzed their metabolomic data with 3 samples in each treatment. In this experiment, 5 samples are taken for each treatment. Small sample sizes are not uncommon in metabolomics, and they present difficulties for using assumptions of normality or application to the central limit theorem.

Dependent Data Structure: In metabolite data, if there is a long chain of reactant and product pathways, one lipid's concentration level change may be associated with all other lipids on the pathway. Therefore, one lipid concentration change in the network might cause a sequence of changes in the pathway or the pathway networks (Steuer et al. 2003). The

fluctuation of concentration in the lipid species can reflect the underlying biological process and it stops at the moment when the researcher takes the sample. The dependence of the lipid data is due to the nature of the underlying pathways. Each sample has 141 variables, some or all of which may be correlated. It is reasonable to assume that the samples themselves are independent of each other.

3.2. Data Manipulation

3.2.1 Centering and Scaling

The following notations will be used for one reactant A and product B in a lipid pair.

n: The sample size in each group.

i: Subscript i = 1, 2 to denote the treatment, 1 as WT and 2 as MT.

j: Subscript j = 1, 2, …, n be sample within treatment.

Before scaling:

x_{ij} : The concentration level for the j^{th} sample in the i^{th} treatment for one lipid.

$\bar{x}_{i\bullet}$: The group mean in the i^{th} treatment for one lipid.

$\bar{x}_{\bullet\bullet}$: The overall mean across two treatment groups for one lipid, $\bar{x}_{\bullet\bullet} = \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n} x_{ij}$.

s : Standard deviation for one lipid across two treatments, $s = \sqrt{\frac{\sum_{i}^{2}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})^2}{2n-1}}$.

x_{Aij} : The concentration of j^{th} sample for lipid A in the i^{th} treatment.

x_{Bij} : The concentration of j^{th} sample for B in the i^{th} treatment.

$\bar{x}_{A\bullet\bullet}$: The mean concentration of A across two treatments.

$\bar{x}_{B\bullet\bullet}$: The mean concentration of B across two treatments.

$\bar{x}_{Ai\bullet}$: The mean concentration of A in the i^{th} treatment.

After scaling:

z_{ij} : The concentration level for the j^{th} sample in the i^{th} treatment for one lipid. Let

$z_{ij} = \frac{x_{ij} - \bar{x}_{\bullet\bullet}}{s}$.

$\bar{z}_{i\bullet}$: The mean concentration in the i^{th} treatment for one lipid and $\bar{z}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} z_{ij}$.

$\bar{z}_{\bullet\bullet}$: The overall mean across two treatment groups for one lipid (note that after scaling, this quantity is exactly zero).

z_{Aij} & z_{Bij} : The concentration level for the j^{th} sample in the i^{th} treatment for reactant A or product B.

$\bar{z}_{Ai\bullet}$: The mean concentration of A in the i^{th} treatment (i.e., $\bar{A}_w$ or $\bar{A}_m$)

$\bar{z}_{Bi\bullet}$: The mean concentration of B in the i^{th} treatment (i.e., $\bar{B}_w$, or $\bar{B}_m$)

In the following part of this analysis, data from the fad2 experiment is used as an illustration for these notations. All other 8 datasets have similar properties. Data for only 4 of 141 lipids are shown in Table 3.1. Concentration levels are given for each lipid in each sample. The unit of the data is *nmol* per mg dry weight. The first 5 samples are from the wild type group (sometimes referred to herein as the 'control group') and the last 5 samples are from the mutant group (sometimes referred to as a 'treatment group'). The same 141 lipid compounds are analyzed in each of the 9 datasets.

Table 3.1: Partial data structure from fad2

Names	WT1	WT2	WT3	WT4	WT5	MT1	MT2	MT3	MT4	MT5
DGDG34_6	2.249	2.175	2.526	1.956	2.212	3.249	2.553	4.058	2.959	2.785
DGDG34_5	0.141	0.103	0.096	0.166	0.061	0.185	0.193	0.074	0.19	0.236
DGDG34_4	0.258	0.194	0.203	0.177	0.226	0.232	0.291	0.316	0.18	0.27
DGDG34_3	4.694	4.329	4.405	3.858	4.026	3.451	3.398	4.164	3.243	3.026

Each lipid concentration has a different mean and a standard deviation when measured across the combined treatment groups. This presents challenges to evaluate reactant-product pairs in a pathway because some lipids are far more abundant in some samples than in others. To put all lipids on the same scale of measurement, their concentrations are centered and scaled by using the standardization formula z_{ij} from the notation defined earlier.

♦ **Proposition 3.1:** Consider a single lipid and denote the concentration by x_{ij} for the j^{th} sample in the i^{th} treatment, where $i = 1, 2$, and $j = 1, 2, \ldots, n$. Then, $|\bar{z}_{i\bullet}| \le \sqrt{1-\frac{1}{2n}}$ for $i = 1, 2$ and $|\bar{z}_{1\bullet} - \bar{z}_{2\bullet}| \le 2\sqrt{1-\frac{1}{2n}}$.

Proof: It is clear that $\bar{z}_{\bullet\bullet} = \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n} z_{ij} = 0$ and equal samples in each group implies $\bar{z}_{1\bullet} = -\bar{z}_{2\bullet}$. So if we focus on $\bar{z}_{1\bullet}$, we have

$$\bar{z}_{1\bullet} = \frac{1}{n}\sum_{j=1}^{n} z_{1j} = \frac{1}{n}\sum_{j=1}^{n}\frac{x_{1j} - \bar{x}_{\bullet\bullet}}{s} = \frac{1}{n}\frac{\sum_{j=1}^{n}(x_{1j} - \bar{x}_{\bullet\bullet})}{s}. \tag{3.1}$$

The numerator of (3.1) can be calculated as

$$\sum_{j=1}^{n}(x_{1j} - \bar{x}_{\bullet\bullet}) = \frac{1}{2}(n\bar{x}_{1\bullet} - n\bar{x}_{2\bullet}) = \frac{n}{2}(\bar{x}_{1\bullet} - \bar{x}_{2\bullet}).$$

By the definition of standard deviation, it can then be converted to

$$\begin{aligned}
(2n-1)s^2 &= \sum_{i=1}^{2}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})^2 = \sum_{j=1}^{n}(x_{1j} - \bar{x}_{\bullet\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{\bullet\bullet})^2 \\
&= \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet} + \bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet} + \bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})^2 \\
&= \sum_{j=1}^{n}\left\{(x_{1j} - \bar{x}_{1\bullet})^2 + (\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})^2 + 2(x_{1j} - \bar{x}_{1\bullet})(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})\right\} \\
&\quad + \sum_{j=1}^{n}\left\{(x_{2j} - \bar{x}_{2\bullet})^2 + (\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})^2 + 2(x_{2j} - \bar{x}_{2\bullet})(\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})\right\} \\
&= \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet})^2 + n(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})^2 + 2(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})\sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet}) \\
&\quad + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet})^2 + n(\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})^2 + 2(\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})\sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet}) \\
&= \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet})^2 + n(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})^2 + n(\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})^2 \\
&= \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet})^2 + n\left(\bar{x}_{1\bullet} - \frac{\bar{x}_{1\bullet} + \bar{x}_{2\bullet}}{2}\right)^2 + n\left(\bar{x}_{2\bullet} - \frac{\bar{x}_{1\bullet} + \bar{x}_{2\bullet}}{2}\right)^2
\end{aligned}$$

then

$$(2n-1)s^2 = \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet})^2 + \frac{n}{2}(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2. \tag{3.2}$$

With (3.2) , we can convert (3.1) into squared treatment mean. Then,

$$\bar{z}_{1\bullet}^2 = \left(\frac{1}{n} \frac{\sum_{j=1}^{n} (x_{1j} - \bar{x}_{\bullet\bullet})}{s} \right)^2 = \frac{(2n-1)}{n^2} \cdot \frac{\left(\frac{n}{2} (\bar{x}_{1\bullet} - \bar{x}_{2\bullet}) \right)^2}{\sum_{j=1}^{n} (x_{1j} - \bar{x}_{1\bullet})^2 + \sum_{j=1}^{n} (x_{2j} - \bar{x}_{2\bullet})^2 + \frac{n}{2} (\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}.$$

For one lipid, let $SSW_1 = \sum_{j=1}^{n} (x_{1j} - \bar{x}_{1\bullet})^2$, and $SSW_2 = \sum_{j=1}^{n} (x_{2j} - \bar{x}_{2\bullet})^2$, then

$$\bar{z}_{1\bullet}^2 = \frac{(2n-1)}{4} \frac{(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}{SSW_1 + SSW_2 + \frac{n}{2} (\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}.$$

For the reciprocal of $\bar{z}_{1\bullet}^2$, $\frac{1}{|\bar{z}_{1\bullet}^2|} = \frac{4}{2n-1} \frac{SSW_1 + SSW_2 + \frac{n}{2} (\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}{(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}$,

i.e., $$\frac{1}{\bar{z}_{1\bullet}^2} = \frac{2n}{2n-1} + \frac{4}{2n-1} \frac{SSW_1 + SSW_2}{(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}. \quad (3.3)$$

Let $\Delta = \frac{4}{2n-1} \frac{SSW_1 + SSW_2}{(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}$, then $\frac{1}{\bar{z}_{1\bullet}^2} = \frac{2n}{(2n-1)} + \Delta$. So $\bar{z}_{1\bullet}^2 = \frac{1}{\frac{2n}{(2n-1)} + \Delta}$.

The max ($\bar{z}_{1\bullet}^2$) should occur when $\Delta \to 0$ which begins to happen when $\bar{x}_{1\bullet} - \bar{x}_{2\bullet} > 0$ and the two within sums of squares, SSW_1 and SSW_2, are close to zero. We have $\bar{z}_{1\bullet}^2 \le \frac{2n-1}{2n}$ or $|\bar{z}_{1\bullet}| \le \sqrt{1 - \frac{1}{2n}}$. Similarly, $|\bar{z}_{2\bullet}| \le \sqrt{1 - \frac{1}{2n}}$. Therefore, result $|\bar{z}_{1\bullet} - \bar{z}_{2\bullet}| \le 2\sqrt{1 - \frac{1}{2n}}$ holds. When n becomes large this upper bound close to 2. ■

After being centered and scaled, all the datasets have mean 0 and standard deviation 1 for each lipid when measured across both the WT and MT groups. In this way, we can compare the arbitrary lipid pairs in a standard unit across all possible pairs. Some noteworthy results follow from the above. According to Proposition 3.1, when there are equal numbers of samples in each treatment group, we have the following results.

1. For the data described here for n = 5, $|\bar{z}_{i\bullet}| \le \sqrt{1 - \frac{1}{10}} = 0.949$. As n gets large, $\max|\bar{z}_{i\bullet}|$ goes to 1.

2. For a lipid pair, A and B, with two dimensional means given by $(\bar{z}_{A1\bullet}, \bar{z}_{B1\bullet})$ and $(\bar{z}_{A2\bullet}, \bar{z}_{B2\bullet})$ for the wild type and mutant, the maximum Euclidian distance between them is 2.684 for $n = 5$ and goes to $\sqrt{8} = 2.828$ as $n \to \infty$.
3. Defining

$$SS_{between} = SSD = (\bar{z}_{A1\bullet} - \bar{z}_{A2\bullet})^2 + (\bar{z}_{B2\bullet} - \bar{z}_{B1\bullet})^2 = SS_{between,A} + SS_{between,B} \text{ and}$$

$$SS_{within} = \sum_{i=1}^{2}\sum_{j=1}^{5}\left[(z_{Aij} - \bar{z}_{Ai\bullet})^2 + (z_{Bij} - \bar{z}_{Bi\bullet})^2\right] = SS_{within,A} + SS_{within,B},$$

   a. $SS_{between} = 0$ implies that the lipid means $|\bar{z}_{1\bullet}| \approx 0$ for each lipid.
   b. If $SS_{within} >> SS_{between}$, both group centers are close to the origin (0, 0).

Another result worth noting is that Pearson's sample correlation coefficient between a pair of lipids is unchanged after scaling the data as described above. This is also true of a population correlation between two variables after centering and scaling the variables by their population mean and standard deviation, respectively. This result is well known and straightforward to show. Table 3.2 shows the data structure after the scaling.

Table 3.2: Illustration of the centered and scaled data

The second and the third columns are the concentration levels. z_{ij} denotes the concentration level with subscript i = 1, 2 to denote the treatment groups and subscript j = 1, 2 ,…, 5 to denote the samples in each group. The last two columns show the mean and standard deviation for each lipid, respectively.

Lipid	WT	MT	Mean	SD
1	$z_{11}, z_{12}, z_{13}, z_{14}, z_{15}$	$z_{21}, z_{22}, z_{23}, z_{24}, z_{25}$	0	1
2	.	.	0	1
.	.	.	.	.
141	$z_{11}, z_{12}, z_{13}, z_{14}, z_{15}$	$z_{21}, z_{22}, z_{23}, z_{24}, z_{25}$	0	1

Figure 3.1 shows the relative position in the WT and MT groups for the same lipid pair before and after scaling. The relative positions of the WT and MT groups remain the same in the two plots.

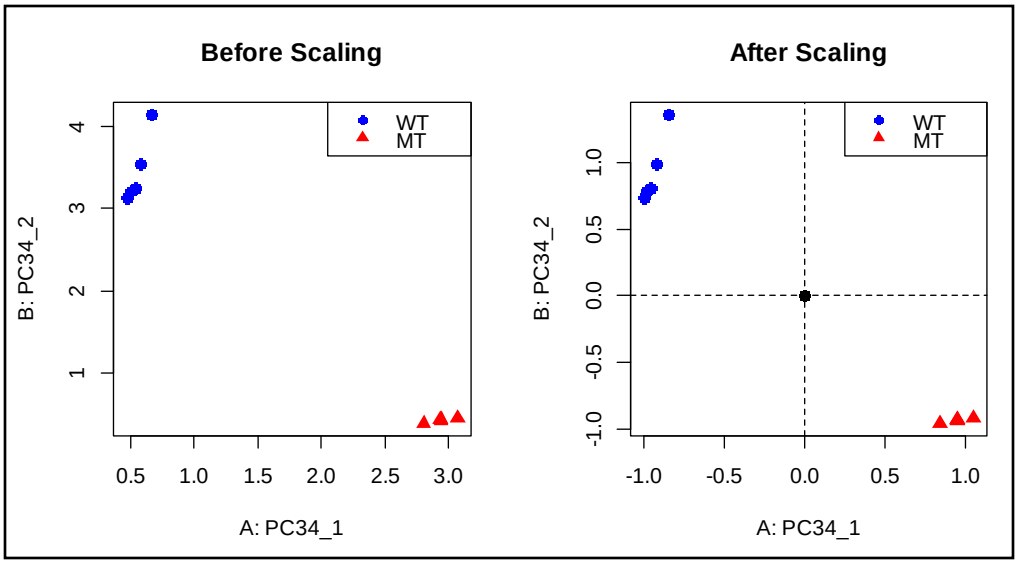

Figure 3.1: Example scatter plots of one lipid pair before and after scaling
Lipid PC34_1 (A, PC34:1, putative reactant) and lipid PC34_2 (B, PC34:2, putative product), which form a lipid pair, are plotted before (left panel) and after scaling (right panel).

After centering and scaling, lipids are paired to determine whether or not the concentration change in the pair follows the scheme $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$. Before screening the lipid pairs according to the scheme, a total of 19740 lipid pairs are created using the first lipid as the reactant A and the second lipid as the product B. Note that each lipid is allowed to be a candidate product or reactant prior to the screening procedure as described next.

3.2.2 Using Variable y to Screen for the Population Lipid Pairs of Interest

Define a variable y, where

$$y = I_{\{\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}\}} + I_{\{\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}\}}.$$

If only $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ is satisfied, then A is a possible reactant. If only $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ is satisfied, then B is a possible product. In both the above cases, $y = 1$. When both $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ hold, A and B are a possible reactant-product pair, and then $y = 2$. Since $y = 2$ gives the results we are looking for, all the arbitrary lipid AB pairs that satisfy the condition $y = 2$ will be used to represent a population of interest in the following sections. Note that y = 0 reflects to the same pair but product and reactant role is reversed, and that $y = 1$ implies that both lipids are either reactants or products but not a reactant-product pair.

3.3. Illustration of Treatment Mean Relationships from Scatter Plots

3.3.1. Patterns in Scatter Plots When y =2 and y = 1

In high dimensional (or any) multiple testing procedures, we usually explore the characteristics of the dataset to summarize some useful test statistics, and then use those statistics to make inferences and draw conclusions to answer the researchers' questions (Westfall 1993). Figure 3.2 shows some scatter plots depicting relationships between two lipids.

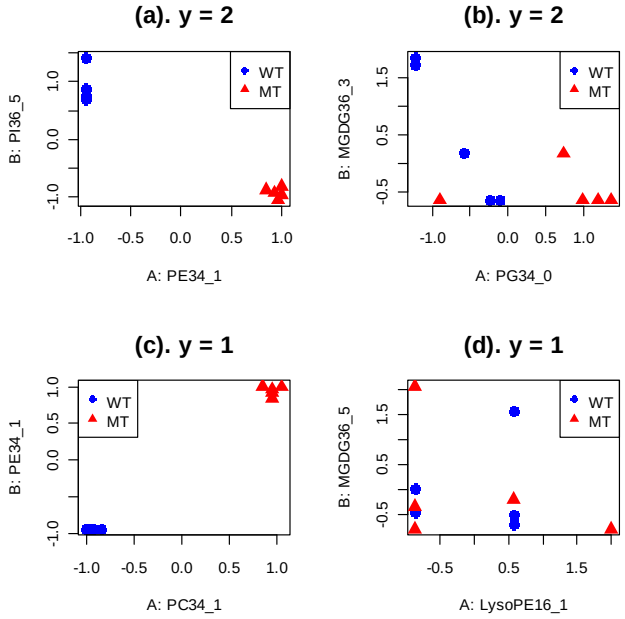

Figure 3.2: Four representative scatter plots from y = 2 and y =1

z_{A1j}, z_{B1j}, z_{A2j} and z_{B2j} stand for the j^{th} sample of a paired lipid in wild type or mutant type for a reactant A or product B. The blue filled circles represent the WT group with (z_{A1j}, z_{B1j}), and the red filled triangles represent the MT with (z_{A2j}, z_{B2j}) from 4 different lipid pairs when y = 2 and y = 1. In (a), PC34_1 is the reactant and PI36_5 is the product in this lipid pair.

We will see that Figure 3.2(a) has the 'ideal' pattern with y = 2 when both the screening scheme conditions $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ are satisfied. The 5 WT data points are on the upper left corner and the 5 MT data points are on the lower right corner. The WT group has clear a separation with the MT group. Figure 3.2(b) is also from y = 2 with the mean relations $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$, but the two groups do not have the same clear separation. The relations of the four means in this pair still satisfy the screening scheme, but the mean differences are not large enough to separate the two groups. Figures 3.2(c) and 3.2(d) are scatter plots for two

different lipid pairs. In both cases, only one of the two conditions, $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$, holds, i.e., $y = 1$. But Figure 3.2(c) shows a clear separation between the two groups. Since the relations of the 4 means for this pair do not satisfy the screening scheme, this pair is not a reactant-product pair candidate. In contrast, Figure 3.2(d) does not show clear separation between the two groups, and this pair will be excluded from further consideration.

3.3.2. Using Biologically AB Pairs as a Criterion for Identifying all AB Pairs

We have a list of reactant and product pairs that are biologically plausible in the biochemical reaction pathways. An important goal is how to determine the biochemical feasibility of a lipid pair. The following features of the biochemical molecular structure can show the attributes from biologically functional lipid pairs (Fan 2010).

- **Possibilities of changing the lipid class.**

For example, the class PC lipid can be catalyzed in a reaction to become a PE class lipid by changing the number of carbons or double bounds in the PC class. It may not be possible to change the MGDG class to the DGDG class, or the PC class to PA class, because the structure of the MGDG class and PC class molecules can not be changed to the DGDG class and PA class by changing either the carbons or the double bounds.

- **Possible changes to a fatty acid tail (acyl chain).**

Within the same lipid class, one lipid can change into another lipid by changing the carbon numbers or by changing the number of double bonds. But in one lipid acyl (tail) chain the double bonds can not be added by more than 3 when the lipid tail has a fixed number of carbons. For example, in PC class, both PC34:1 and PC34:5 are from the PC lipid class with 34 carbons and 1 double bond in PC34:1, and 34 carbons with 5 double bonds in PC34:5. It is not possible to change PC34:1 to PC34:5 because the number of double bonds that can be added to one acyl tail is from 0 to 3. While in the same class of lipid with multiple acyl tails, the PC34:1 can be changed to PC34:4 and PC34:1 to PC40:4, but the head group PC can not be changed at the same time.

- **Extending or shortening a fatty acid chain.**

One lipid can be converted to another by changing the double bond numbers in the acyl chain. For example, to change a X to lyso X, 0 to 3 of the double bonds should be removed from one acyl chain.

These biological reactant and product pairs are used as a standard to compare with all other arbitrary lipid pairs. Figure 3.3 shows the scatter plot characteristics of the biologically functional reactant-product pairs in the dataset fad2. All other 8 datasets' biologically functional lipid pairs have similar properties with those in fad2. There are in total 17 known biologically functional lipid pairs in the fad2 dataset. The rest scatter plots from the biologically functional ones in other datasets are all similar to those in Figure 3.3. All the biologically functional lipid pairs scatter plots have similar patterns: WT is at the upper left corner, MT is at the lower right corner. Their concentration relationships satisfy $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ in all biologically functional lipid pairs with a large mean difference in each lipid. Figure 3.4, summarizes scatter plots in all the biologically pairs.

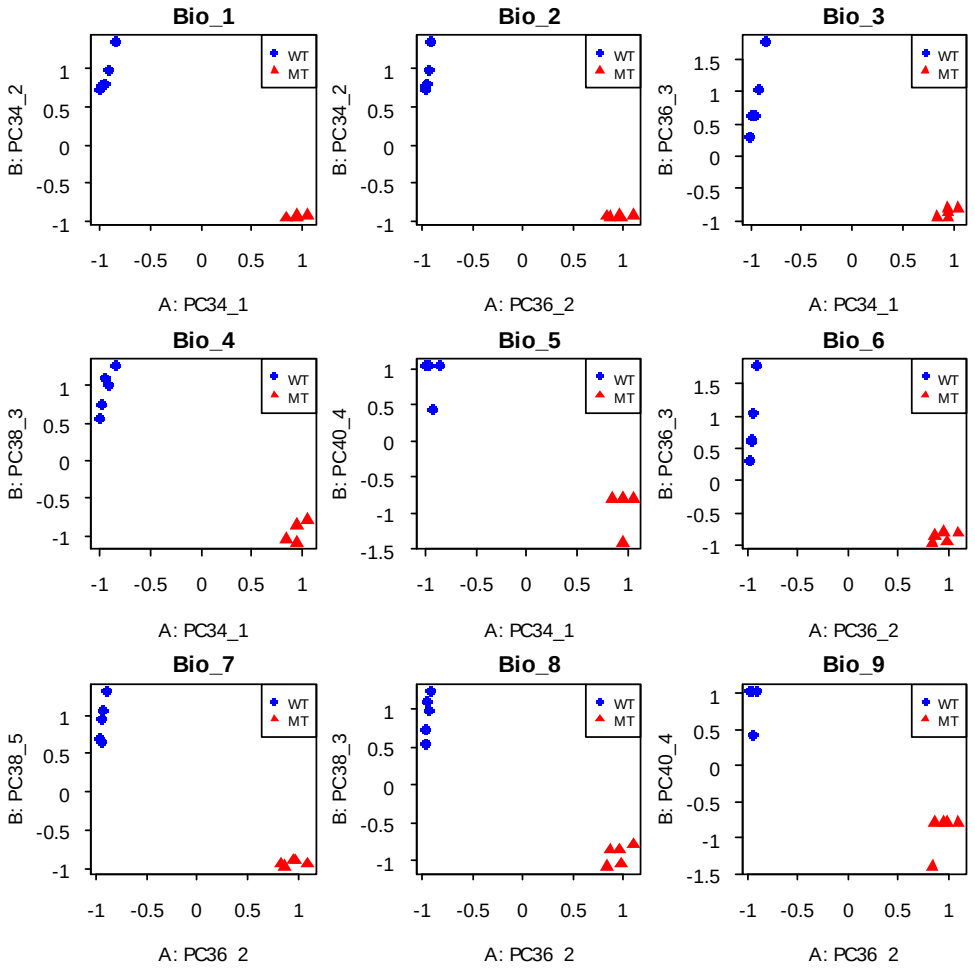

Figure 3.3: Scatter plots of some biologically functional lipid pairs in *fad2*

In each panel, the 5 blue circles represent the WT group with coordinates (z_{A1j}, z_{B1j}) and the 5 red triangles stand for the MT group with coordinates (z_{A2j}, z_{B2j}). The x-axis is the concentration of the reactant A and the y-axis is the concentration of the product B.

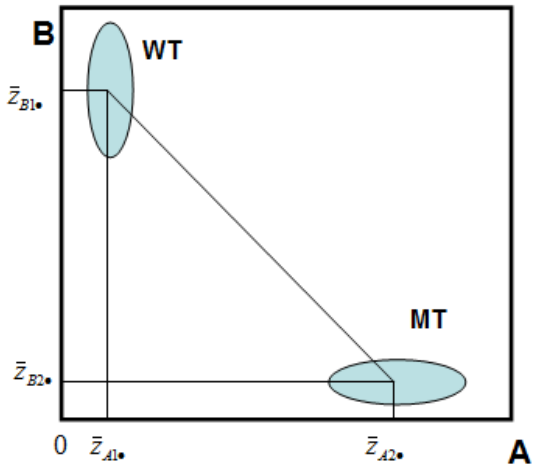

Figure 3.4: A typical reactant and product candidate pair after scaling.
The two ovals represent wild type group (WT) and mutant type group (MT), respectively. While $\bar{z}_{A1\bullet}$ and $\bar{z}_{B1\bullet}$ stand for wild type reactant and product group means, $\bar{z}_{A2\bullet}$ and $\bar{z}_{B2\bullet}$ stand for mutant type reactant and product group means, respectively. The three points, $(\bar{z}_{A2\bullet}, \bar{z}_{B2\bullet})$, $(\bar{z}_{A1\bullet}, \bar{z}_{B1\bullet})$ and $(\bar{z}_{A1\bullet}, \bar{z}_{B2\bullet})$, form a right triangle.

The correlation analysis technique of Fukushima et al. (2011) was discussed in chapter 2. The next section explores the application of this method to the data used herein.

3.4. Can Correlation Analysis Work in our Lipid Experiments?

We explore the 9 datasets in lipid experiments by using the correlation data analysis procedure by Fukushima et al. (2011). The notations from Fukushima et al. (2011) are adapted by using X and Y as a lipid pair without discrimination of the role of reactant A and product B. Define $r_1 = r_{X_w Y_w}$ as the correlation between X_w and Y_w in the WT and $r_2 = r_{X_m Y_m}$ as the correlation between X_m and Y_m in the MT.

Fukushima et al. were interested in the correlations differences $r_1 - r_2$ in metabolite pairs between roots and aerial tissues. In this correlation analysis, the single correlations in each treatment are ignored, because our research interest is to find the significant correlation differences, $r_1 - r_2$, between the wild type and the mutant treatments which can reflect the mutation effect in lipid pairs. Therefore, the Z test from Fisher's transformation on the correlation difference $r_1 - r_2$ is used to evaluate the significance of mutation effect for each lipid pair from test statistic (2.4) at *lfdr* level of 0.05. Tests of the single correlations are not of primary interest; however, we still evaluate the significance of single correlations so as to assess

whether the number of significant correlations is different across treatment conditions. The t-test in (2.2) is used to evaluate the significance of $r_1 = r_{X_w Y_w}$ and $r_2 = r_{X_m Y_m}$ at *lfdr* level of 0.05 in a test of whether there is evidence that the two lipids are associated or not (i.e., a two tailed test of a null hypothesis of no association). Spearman's correlation and the Pearson's correlation are utilized separately to get results. Since the biologically functional lipid pairs are the criterion to see if the analysis results are reasonable or not, using the biologically functional lipid pairs to judge the correlation analysis results will be the strategy.

Table 3.3 shows the analysis results using Spearman's correlation. One can see that the numbers of significant correlations in WT are significantly larger than those in MT by comparing the first row and the second row. But in the last row, Spearman's correlation does not find any lipid pair that has a significant correlation difference. Since Spearman's correlation uses ranks to get the results, when we have only 5 samples, a small sample size and many ties among ranks can create too much discreteness in the randomization distribution and low power for detecting significant results.

Since Spearman's correlation may not be a suitable metric for dependence as it was in Fukushima et al. (2011), Pearson's correlation was also used in the correlation analysis. The results are shown in Table 3.4.

One can see that Pearson's correlation analysis detects more significant results than use of Spearman's correlation analysis. However, are those results reliable? Can the correlation analysis reflect the main characteristics of the lipid data via the biologically functional lipid pairs? To check Pearson's correlation analysis results, the lipid pairs with significant correlation differences are matched with the biologically functional lipid pairs in 9 lipid datasets in Table 3.5.

Table 3.3: Number of significant Spearman's correlations in the 9 experiments

The first row stands for the number of significant correlations, $r_1 = r_{X_w Y_w}$, from the WT for each experiment. Row 2 stands for the number of significant correlations, $r_2 = r_{X_m Y_m}$, from the MT. Row 3 shows the number of significant correlation differences.

Correlations	fad2	fad3	fad4	fad5	fad6	fad7	sfd1	sfd2	sfd3
$r_1 = r_{X_w Y_w}$	177	177	177	177	177	177	149	34	149
$r_2 = r_{X_m Y_m}$	63	8	2	6	5	2	6	98	2
$r_1 - r_2$	0	0	0	0	0	0	0	0	0

Table 3.4: Number of significant Pearson's correlations in the 9 experiments
The first row stands for the number of significant correlations, $r_1 = r_{X_w Y_w}$, from the WT. Row 2 stands for the number of significant correlations, $r_2 = r_{X_m Y_m}$, from the MT. Row 3 shows the number of pairs with significant correlation differences.

Correlations	fad2	fad3	fad4	fad5	fad6	fad7	sfd1	sfd2	sfd3
$r_1 = r_{X_w Y_w}$	3	3	0	3	3	3	13	71	13
$r_2 = r_{X_m Y_m}$	12	6	0	9	6	0	1121	28	11
$r_1 - r_2$	16	12	0	0	10	15	18	0	0

Table 3.5: Matching number between the biologically functional lipid pairs and the pairs with significant correlation differences from Pearson's correlation analysis
The match is performed in all the 9 experiments. First row shows the number of biologically functional lipid pairs that are found from the significant results from the correlation differences in row 3 of Table 3.4. For comparison, the total number of biologically functional lipid pairs given by experimental biologists is shown in the second row for each experiment. The last row stands for the total number of arbitrary lipid pairs in each dataset.

	fad2	fad3	fad4	fad5	fad6	fad7	sfd1	sfd2	sfd3
Biological pair matched	0	0	0	0	0	0	0	0	0
Total biologically lipid pairs	17	23	7	16	8	14	2	5	7
Total number of pairs	9180	8001	8256	8385	9045	8646	8911	7381	9180

The first row of Table 3.5 shows that no biologically functional lipid pairs are matched with the results in any of the 9 datasets. For example, even though there are 9180 arbitrary lipid pairs in fad2 and 17 biologically functional lipid pairs, none of the biologically functional lipid pairs can be matched with any of the 16 lipid pairs shown in the last row of Table 3.4. Therefore, Pearson's correlation analysis does not seem helpful in finding the significant reactant and product lipid pairs that have similar characteristics with the biologically functional lipid pairs.

Figure 3.5 shows the significant mutation effect using Pearson's correlation results from Table 3.4. One can see that in the first panel of WT, the tallest bar is sfd2. This means that more lipids in sfd2 in the WT group are correlated. In panel MT group, the tallest bar is the sfd1. It means that the more lipids in sfd1 in the MT group are correlated. While in the correlation differences in the third panel, there are not many lipid pairs that are affected significantly.

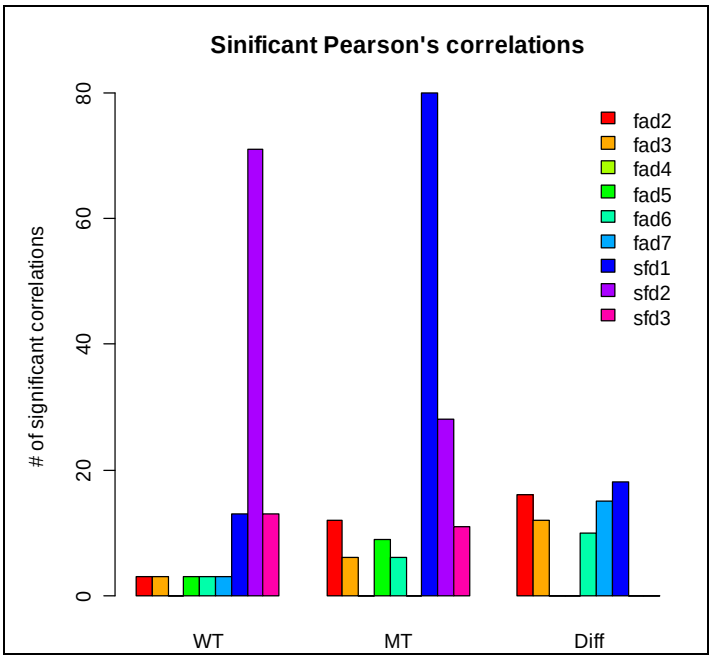

Figure 3.5: The number of significant correlations comparison in all 9 datasets
This graph is produced from Table 3.4 to visualize the mutation effect by using Pearson's correlation. The same color stands for the same data in each of the three group comparisons.

In conclusion, Pearson's correlation analysis can find some significant lipid pairs according to the linear association between the concentration levels in each group. But the issue is that the correlation within WT or MT is not a good metric to describe a biologically functional reactant-product pair affected by the mutation. To achieve our research goal, a novel statistical data analysis is needed.

3.5. Some Numerical Facts from the Exploratory Data Analysis

Figure 3.6 illustrates the relations among the treatment means in a candidate A-B pair in the two dimensional space from Result 3. For A, $0 < \bar{z}_{A2\bullet} < 0.949$ and $-0.949 < \bar{z}_{A1\bullet} < 0$. Similarly for B, $0 < \bar{z}_{B1\bullet} < 0.949$ and $-0.949 < \bar{z}_{B2\bullet} < 0$. The positions of the WT and MT groups shown in this figure stand for the extreme cases when $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ hold. Since the treatment groups can have many different positions than those in this plot, the group centers, $(\bar{z}_{A1\bullet}, \bar{z}_{B1\bullet})$ and $(\bar{z}_{A2\bullet}, \bar{z}_{B2\bullet})$ can be in anywhere within the shaded 2 × 2 square area. The border of the square represents the limits of the A-B lipid pair treatment means.

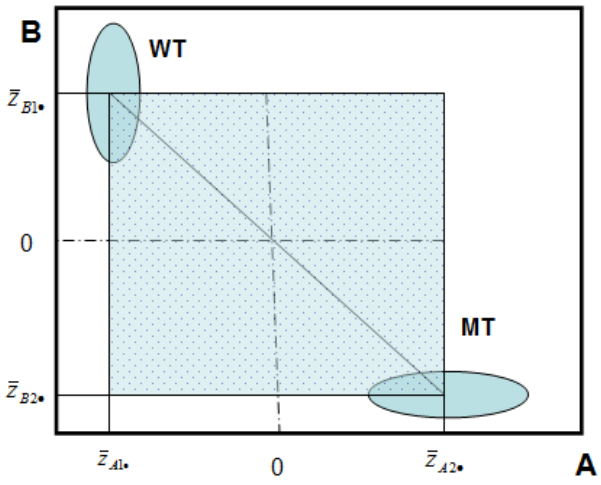

Figure 3.6: The relations of the treatment means for an AB lipid pair
The shaded square represents the possible positions for $(\bar{z}_{A2\bullet}, \bar{z}_{B2\bullet})$ and $(\bar{z}_{A1\bullet}, \bar{z}_{B1\bullet})$ which are the wild type and mutant type groups for all arbitrary lipid pairs after scaling, respectively.

Figure 3.7(a) shows the scatter plot of WT and MT for all the 4623 lipid pairs with $y = 2$, (recall that $y = 2$ if the lipid pair means satisfy the conditions $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$). The WT groups (gray) would appear at the top-left corner and the MT groups (light blue) would be located at the bottom-right corner. The blue points and the red points are the WT and MT groups from 17 biologically functional lipid pairs, respectively. Since the 135^{o} reference blue line crosses the centers of the biologically functional pair WT and MT groups, this position is a chosen for a possible reactant and product in a lipid pair. This position should be 135 º angles to the x-axis, and the WT and MT groups should be separated with a large inter-group distance.

Figure 3.7(b) shows that within the 4623 lipid pairs with $y = 2$, the WT treatment group centers $(\bar{z}_{A1\bullet}, \bar{z}_{B1\bullet})$ fall in the 0.949 by 0.949 black square area, and the MT treatment group centers $(\bar{z}_{A2\bullet}, \bar{z}_{B2\bullet})$ fall in the 0.949 by 0.949 yellow square area. According to Proposition 3.1 and its results, the treatment centers $(\bar{z}_{A1\bullet}, \bar{z}_{B1\bullet})$ in the black square and $(\bar{z}_{A2\bullet}, \bar{z}_{B2\bullet})$ in the yellow square for any lipid pair are symmetric about (0, 0). Since the biologically functional lipid pairs are all from these lipid pairs, they are a useful source from which the significant results are drawn. Several statistics can be summarized from these lipid pairs to reflect the mean relationships.

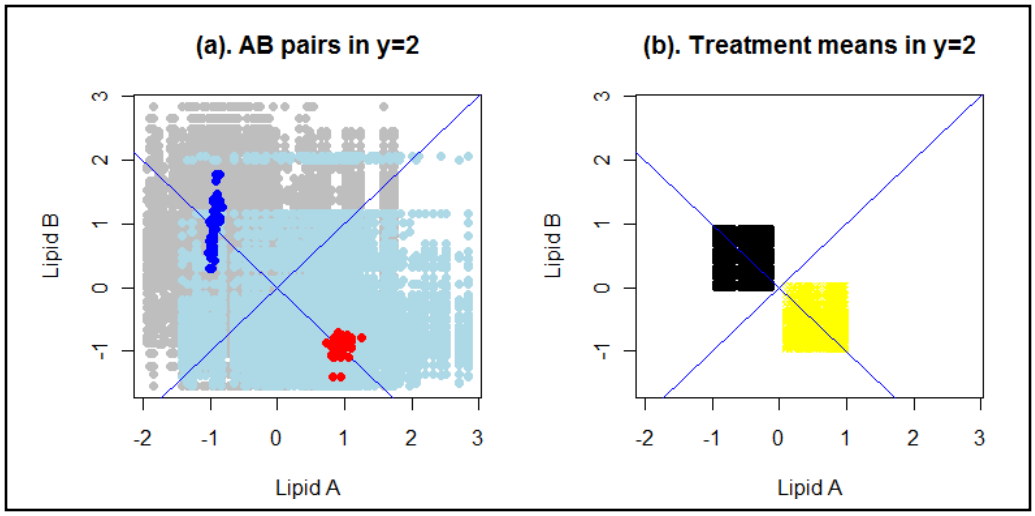

Figure 3.7: The scatter plot and the treatment mean plot when y =2 from dataset fad2
(a) The light gray (WT) and the light blue (MT) data points are the scatter plot for all the 4623 possible lipid pairs in fad2 which satisfy the screening scheme relations $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ or equivalently $y = 2$. The blue (WT) and red (MT) data points are the scatter plot for all the 17 biologically functional lipid pairs. The diagonal blue lines are reference lines with 45º and 135º angles relative to the x-axis. **(b)** The treatment means from the lipid pairs with $y = 2$. The yellow square represents the area in which the MT group means fall and the black square is the area in which the WT group means fall.

3.6. Three Summary Statistics

Three summary test statistics are generated according to the exploratory data analysis. They are tg ratio, SSD and R statistics. The distributions of the test statistics are shown in Figure 3.8. All statistics can be applied to each lipid reactant-product pair in the scaled data.

- **Statistic 1: tg ratio**

This tg ratio is the ratio of lipid product B group mean difference to the reactant A group mean difference which can be defined as

$$tg = \frac{\bar{z}_{B1\bullet} - \bar{z}_{B2\bullet}}{\bar{z}_{A2\bullet} - \bar{z}_{A1\bullet}}, \qquad (3.4)$$

where means $\bar{z}_{ij\bullet}$ are defined early in this chapter. The ideal position for the two dimensional groups WT and MT is when they form a 135 degree angle with the x-axis, which leads to tg = 1. The red lines in the tg ratio distribution in Figure 3.8(a) shows the biologically functional lipid pairs with tg ratio values that are all close to 1.

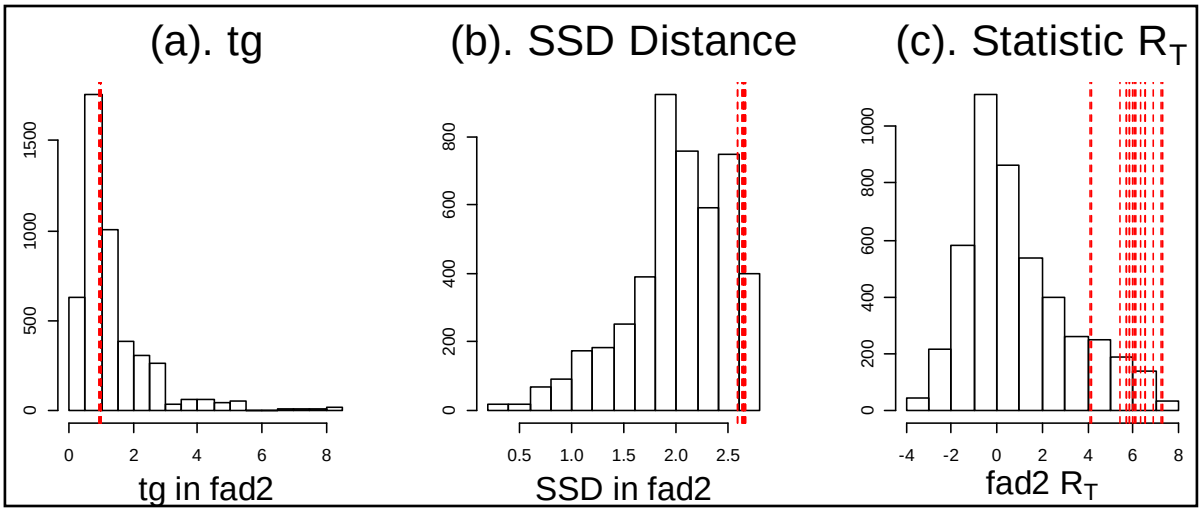

Figure 3.8: The distributions of the test statistics.
The red dashed vertical lines stand for the biological lipid pair statistics which are matched with the three statistics in dataset fad2 from lipid pairs with $y = 2$.

- **Statistic 2: SSD**

SSD is a squared distance between the two group centers as defined early in the Result 2 with the form

$$SSD = SS_{between} = (\bar{z}_{A1\bullet} - \bar{z}_{A2\bullet})^2 + (\bar{z}_{B2\bullet} - \bar{z}_{B1\bullet})^2 \qquad (3.5)$$

Large SSD, or inter-group distance gives results shown in Figure 3.8(b).

- **Statistic 3: R_T statistic**

The lipid pair with tg = 1 or the angle close to 135° between the line formed by the WT and MT type mean centers and the positive x-axis will reflect the true candidate biological pairs in this data analysis. Raamsdonk et al. (2001) used a similar measurement co-response coefficient Ω as a ratio of the log concentration level change in FANCY to illustrate the best position with 45° or −45° that is analogous to tg here. Raamsdonk et al. already showed that this co-response may go to infinity when the reference metabolite mean concentration change is very small (see equation (2.1)), making the evaluation impossible. Using tg ratio has the same disadvantage. Since the tg ratio is also a ratio statistic, it can be very large or close to infinity (or can be very small and close to 0) when the angle is close to 90° (or 0°). Therefore, if used alone, this measurement does not capture all characteristics of potential reactant-product pair. Similarly, the SSD statistic also has some disadvantages. If the inter-group distance is very large, but the

angle between the groups is much different from 135º, then the result may not be real. This means that using SSD alone may also lead to false discoveries.

Therefore, a statistic using both tg and SSD at the same time is proposed. This combined statistic, called R, can eliminate the respective disadvantages of tg and SSD while keeping their advantages. The statistic R is defined as

$$R = (tg - 1)^2 + (SSD - \max(SSD))^2, \qquad (3.6)$$

where tg is the tg ratio as defined in (3.4), and SSD is defined in (3.5). The value $\max(SSD)$ is set to 2.684 from the theoretical maximum SSD from the Result 2 when sample size is 5. This ensures that the value of R will never be exactly zero. As we will see later, for interpretability purposes, the R statistics in (3.6) will be transformed by logarithm. Smallest R values are the lipid pairs that are significantly affected by the mutation.

In Figure 3.9, the data points (bio.tg, bio.SSD) from the 17 different biologically functional lipid pairs indicate that the most interesting pairs are close to the peak (1, 2,684). For convenience small values of R are converted to large values of –log(R) to perform an upper tail test rather than a lower tail test. Figure 3.8(c) shows the distribution of the transformed R statistic using $R_T = -\log(R)$. The biologically functional pair's R_T statistics (red lines) shows that the larger values of R_T will reflect the most interesting results.

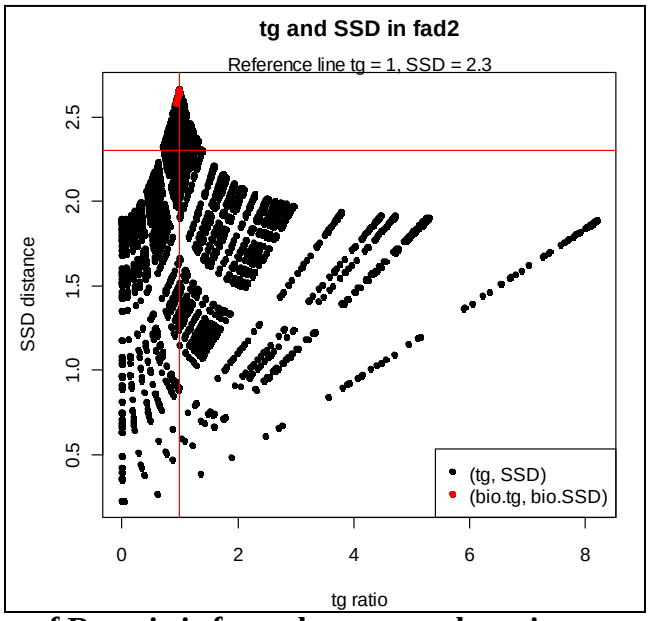

Figure 3.9: Illustration of R statistic from the scatter plot using tg and SSD in dataset fad2 The vertical red line shows tg = 1. The horizontal red line shows an arbitrary cutoff point at SSD = 2.3.

In Figure 3.9, the black points are the (tg, SSD) coordinates for each lipid pair. The red points at the peak are the biologically functional lipid pairs. The peak area contains the most interesting lipid pairs with large SSD values and tg close to 1. Note that in Figure 3.9, the scatter plot of tg and SSD preserves some regular patterns because tg and SSD are all functions of the means, $\bar{z}_{Ai\bullet}$ and $\bar{z}_{Bi\bullet}$ as shown previously.

In conclusion, the three statistics are derived from the exploration of the data according to the screening scheme in Figure 1.5. From the above analysis, we can see that the three statistics can reflect the data characteristics for lipid pairs that are biologically functional reactant-product pairs whose reaction is modified by the mutation in the organism. They can be employed separately or combined as a whole. In the next chapter, I will introduce a bootstrap technique to determine a null distribution of each of the three statistics. This will be one null distribution and, as we will see, it may not be the ideal null distribution since the null hypothesis under which it is derived may be too restrictive. More discussion of this interesting dilemma follows.

Chapter 4 - A Parametric Bootstrap Null Distribution

4.1. Bootstrap Algorithm

Some literature on fitting a null distribution for the mixture model approach was reviewed in chapter 2. In chapter 3 three statistics tg, SSD, and R_T were proposed. In this chapter, a bootstrap procedure is used to establish an approach for getting a null distribution for the three statistics separately using a well-known parametric distribution. Then the parametric bootstrap null (PBN) distribution will be fit to the real data to find a list of discoveries.

Following the screening scheme in Figure 1.5, after scaling, let multivariate WT concentration data be $z_{11}, z_{12}, \ldots, z_{1n} \xrightarrow{\text{iid}} F$, and mutant data be $z_{21}, z_{22}, \ldots, z_{2n} \xrightarrow{\text{iid}} G$. F and G denote the distributions from which the WT and MT data come. z_{ij} denotes the concentration level for the j^{th} sample in the i^{th} treatment groups. The index, i = 1, 2, represents the WT and MT treatments, and the other index, j = 1, …, n, stands for the sample in the i^{th} treatment. Assuming that there is no mutation effect, the null hypothesis $F = G$ should be satisfied for each lipid species in the 141 lipids in the dataset. This will be the assumption under which the bootstrap null distribution is generated in the following. Some relevant notation for later is the following: let μ_{Aw}, μ_{Bw}, μ_{Am}, and μ_{Bm} be the population means for the reactant A in the wild type, product B in the wild type, reactant A in the mutant type, and reactant B in the mutant type treatments, respectively. The statement $F = G$ can be expressed in the following null hypothesis,

$$\begin{aligned} H_0 &: F = G \\ H_A &: F \neq G \end{aligned} \tag{4.1}$$

Note that the above null hypothesis is quite restrictive in saying that the MT and WT data are from exactly the same high-dimensional distribution. Nevertheless, such a null is not uncommon and is the null hypothesis under which some randomization tests and bootstrap tests are carried out. The null distributions of tg^*, SSD^*, and R_T^* will be generated under the null hypothesis in (4.1) from the lipid pairs that satisfy the screening scheme $\bar{z}^*_{A1\bullet} < \bar{z}^*_{A2\bullet}$ and $\bar{z}^*_{B1\bullet} > \bar{z}^*_{B2\bullet}$, or $y^* = 2$. In this section, all the notations used in the bootstrap will be similar with those in the real data except that the statistics from the bootstrap are marked with a star (*). Figure 4.1 is a schematic of the bootstrap procedure.

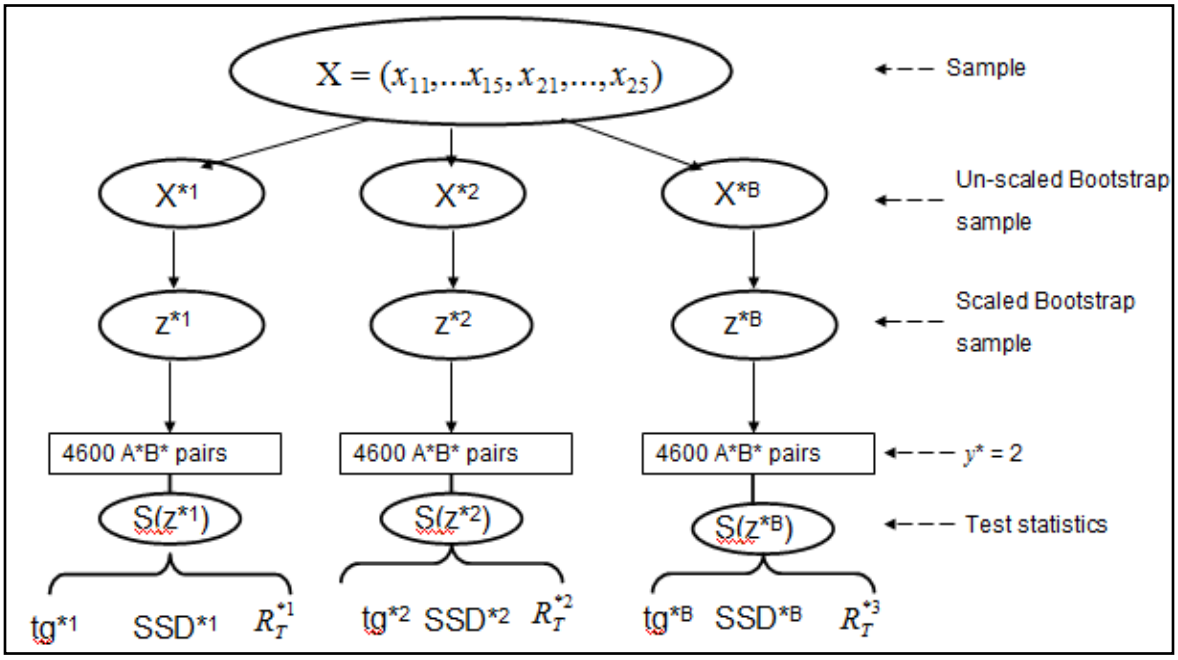

Figure 4.1: Schematic of the bootstrap procedure for generating the three test statistics from each bootstrap sample

Dataset X is the original sample with 141 rows (lipids) and 10 columns (5 WT and 5 MT samples). B bootstrap samples $X^{*1},\ldots, X^{*B}$ are generated from the original data. Center and scale the B bootstrap samples to get the scaled bootstrap samples $z^{*1},\ldots, z^{*B}$. Approximately $K = 4600$ lipid pairs which satisfy the screening scheme $\bar{z}^*_{A1\bullet} < \bar{z}^*_{A2\bullet}$ and $\bar{z}^*_{B1\bullet} > \bar{z}^*_{B2\bullet}$ are generated from the scaled bootstrap samples. 4600 statistics values are computed for each tg^*, SSD^* and R_T^*, respectively, from the scaled bootstrap samples. Using tg^* as an example, if B = 200, a total of 4600×200 tg^* values are generated from the 200 bootstrap samples.

Let B be the number of bootstrap samples. Figure 4.1 shows the flow from the un-scaled sample $X = (x_{11},...,x_{15},x_{21},...,x_{25})$ data to the b^{th} bootstrap samples X^{*b} to the scaled b^{th} bootstrap samples Z^{*b}. The screening scheme is applied to each lipid pair from the scaled bootstrap sample to obtain, say K, test statistics for each statistic tg^*, SSD^*, and R_T^* separately in the b^{th} bootstrap loops. Note that the number K will vary from sample to sample in the bootstrap sampling procedure because of the screening for pairs for which $y^* = 2$. Typically this number is approximately $K = 4600$ and, for convenience, we use the number 4600 in what follows as the technique is described.

Bootstrap Algorithm

1. Let the sample be a matrix $X = (x_{11},...x_{15}, x_{21},..., x_{25})$ with the 141 lipid species representing the rows and with 5 columns for WT samples and 5 columns for MT samples.
2. Let B = 200, and $b = 1, 2,..., B$ to denote the b[th] bootstrap sample.
3. Take random sample from 1 to 10 with replacement to get a vector with 10 integer values. Generate the b[th] bootstrap sample $\mathrm{X}^{*b} = (x_{11}^*,...x_{15}^*, x_{21}^*,..., x_{25}^*)$ with 141 rows and the 10 random numbers as the columns. The first 5 columns are used as the WT and last 5 columns as the MT groups.
4. In the bootstrap sample, reduce (delete) the number of lipid species in each row when either the concentrations are all zero values or the standard deviation is 0 across the two treatment groups for that lipid.
5. Obtain centered and scaled samples $\mathrm{Z}^{*b} = (z_{11}^*,...z_{15}^*, z_{21}^*,..., z_{25}^*)$ by using $z_{ij}^* = \frac{x_{ij}^* - \bar{x}_{\bullet\bullet}^*}{s^*}$ to get the mean 0 and standard deviation 1 for each lipid species. x_{ij}^* and z_{ij}^* denote the bootstrap data before and after scaling.
6. Pair any reactant A^* with any product B^* to get approximately $2\binom{141}{2} = 19740$ lipid pairs $\mathrm{A}^* \rightarrow \mathrm{B}^*$.
7. Screen the relationships of $\bar{z}_{A1\bullet}^* < \bar{z}_{A2\bullet}^*$ and $\bar{z}_{B1\bullet}^* > \bar{z}_{B2\bullet}^*$ according to the value $y^* = 2$ for any arbitrary lipid pairs. Pick out approximately K = 4600 lipid pairs with $y^* = 2$.
8. Calculate the three statistics, tg^*, SSD^*, and R_T^* for each lipid pair $\mathrm{A}^*\mathrm{B}^*$ from the scaled bootstrap sample z^{*b} in the $y^* = 2$ subpopulation. The test statistics are vectors with about 4600 elements from each of the b[th] bootstrap samples,
   i.e., $\mathrm{tg}^{*b} = (tg^{*1}, tg^{*2},..., tg^{*4600})$, $\mathrm{SSD}^{*b} = (SSD^{*1}, SSD^{*2},..., SSD^{*4600})$,
   $\mathrm{R}_\mathrm{T}^{*b} = (\mathrm{R}_\mathrm{T}^{*1}, \mathrm{R}_\mathrm{T}^{*2},..., \mathrm{R}_\mathrm{T}^{*4600})$. The three statistics are defined as in equations (3.4), (3.5) and (3.6) with the forms

$$\mathrm{tg}^* = \frac{\bar{z}_{B*1\bullet}^* - \bar{z}_{B*2\bullet}^*}{\bar{z}_{A*2\bullet}^* - \bar{z}_{A*1\bullet}^*},$$

$$\mathrm{SSD}^* = (\bar{z}^*_{A^*1\bullet} - \bar{z}^*_{A^*2\bullet})^2 + (\bar{z}^*_{B^*2\bullet} - \bar{z}^*_{B^*1\bullet})^2,$$

$$\mathrm{R}^* = (tg^* - 1)^2 + (SSD^* - 2.828))^2$$

$$\mathrm{R}^*_{\mathrm{T}} = -\log(R^*)$$

9. Use the (≈ 4600) bootstrap statistics to determine the null distribution for the three statistics, tg^*, SSD^*, and R^*_T, respectively.

4600 bootstrap statistics from the b^{th} bootstrap sample can form a null distribution in one bootstrap loop. B=200 null distributions are generated from 200 bootstrap samples. The question is how to summarize the 200 bootstrap null distributions into one null distribution for the three statistics tg, SSD and R^*_T, respectively.

4.2 Bootstrap Null Distribution

4.2.1. Choose Distribution Class for the Null Distribution

Once the null distribution of the statistic is determined in one bootstrap sample, the procedure in Figure 4.2 is adopted to obtain one null distribution for each of the statistics, tg, SSD and R^*_T, in the b^{th} bootstrap sample. Note that Figure 4.2 is the procedure only for one bootstrap loop. The vector $\mathrm{S}^{*b} = [s^{*1}, s^{*2}, ...s^{*4600}]^b$ is used to denote a vector of statistics from the b^{th} bootstrap sample $\mathrm{Z}^{*\mathrm{b}}$ with 4600 elements, and $\mathrm{S}^{*\mathrm{b}}$ stands for any of the three statistics: tg^*, SSD^*, and R^*_T.

As shown in Figure 4.2, Maximum Likelihood Estimation (MLE) is used to fit the b^{th} bootstrap statistics $\mathrm{S}^{*b} = [s^{*1}, s^{*2}, ...s^{*4600}]^b$ to each of four well-known distributions, the Exponential, Gamma, Log-normal and Weibull distributions. The statistical software R package "fitdistrplus" is utilized for the model fitting and parameter estimation procedures. In addition to the parameter estimates (MLEs), the Kolmogorov-Smirnov (K-S) test statistic D, AIC and BIC are extracted from the estimation procedure for the four models. Those estimates will be used to compare the four models to determine which type of model is the best candidate null distribution resulting from the 200 bootstrap samples.

The K-S test statistic D is the maximum vertical distance from the empirical CDF to the fitted parametric distribution. AIC is defined as $2k - 2\ln(L)$, where k is the number of the parameters in the model, and L is the maximized value of the likelihood function for the

estimated model. The preferred model has the smallest AIC, i.e. the model that has maximized log-likelihood value and has smallest number of parameters. BIC is closely related to AIC, but the penalty term (i.e. $k\ln(n)$, where n is the sample size and k is the number of the parameters) is more stringent than that of AIC.

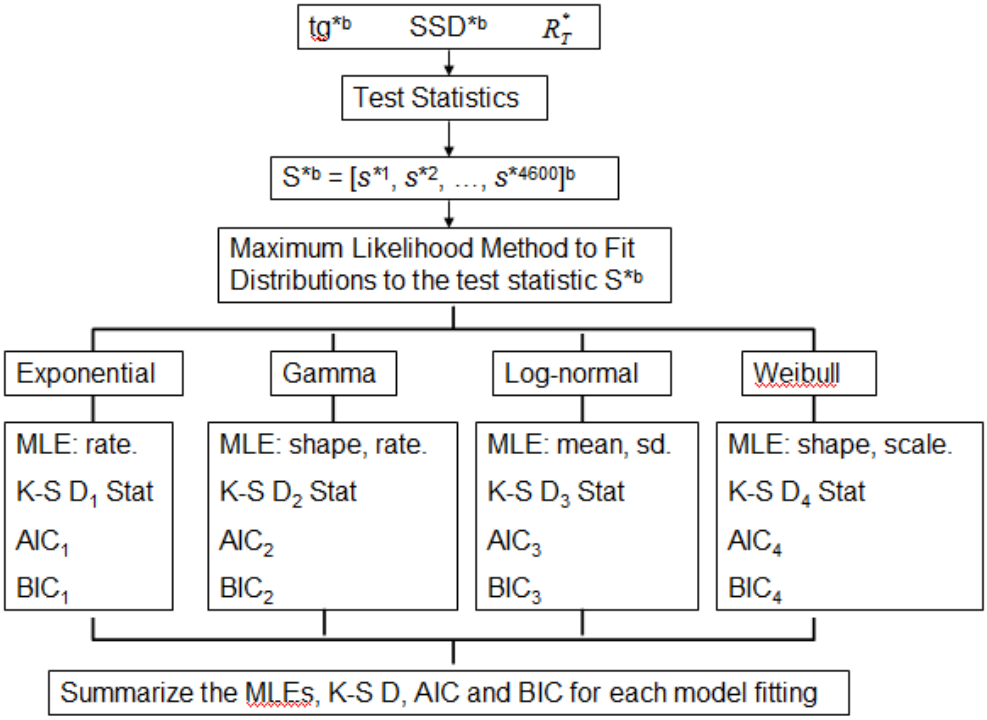

Figure 4.2: Flow chart for generating one null distribution of the statistics using the maximum likelihood estimation procedure in the bth bootstrap sample.

The above chart shows the null distribution generation for the test statistics, tg*, SSD* and R_T^*, from one bootstrap sample $Z^{*b} = (z_{11}^*,...z_{15}^*, z_{21}^*,..., z_{25}^*)$. Four well-known distributions, Exponential, Gamma, Lognormal and Weibull, are used to fit the bth bootstrap test statistics separately. For example, for tg*, tg$^{*b} = (tg^{*1}, tg^{*2},..., tg^{*4600})$. MLE, Kolmogorov-Smirnov (K-S) test statistic D, AIC and BIC are extracted from each model fitting.

The ideas will be illustrated in the context of using the K-S test statistic D from each bootstrap sample to assess the goodness of fit of the bootstrap statistics (i.e., tg, SSD, and R_T^*) to the candidate parametric distribution. Here, the p-values from the K-S test are not used for two key reasons. First, the null hypothesis in a K-S test is that the candidate distribution is the correct one, and the K-S test statistic seeks evidence that the alternative hypothesis (the candidate distribution is not correct) is correct. Very large sample sizes are in play for each K-S test (a

sample of about 4600 observations) so every null hypothesis is likely to be rejected for even very minor departures from the null hypothesis. These departures may not be of any practical importance as long as the candidate distribution is effective in modeling the patterns that are of interest in the null distribution of the test statistics. Second, the 4600 or so observations used in a K-S test are not likely to be independent, making interpretation of p-values questionable. Thus a following procedure is used that utilizes the K-S D statistic as a metric but using the bootstrap to delineate its sampling distribution. The candidate distribution with the smallest K-S D statistic across bootstrap samples is interpreted to be a better candidate than the others.

The procedure in Figure 4.2 is repeated B = 200 times to give 200 bootstrap samples. Suppose 200 null distributions for the statistics are generated. To summarize the 200 null distributions into one final null distribution, all the quantities, K-S D statistic, as well as the AIC and BIC from 200 bootstrap samples are used to determine the preferred candidate distribution to represent the null distribution of the statistics, tg, SSD, and R_T^*. The following details use of the K-S D statistic. The use of AIC and BIC for evaluating candidate distributions is similar.

Minimum K-S Test Statistic D Criteria: The Kolmogorov-Smirnov test (K-S test) determines if two datasets differ significantly. A smaller value of K-S statistic D represents a smaller maximum vertical distance from the empirical data distribution to the fitted parametric distribution. It is a measure of the goodness-of-fit for the 4 candidate distributions to the test statistic $\mathrm{S}^{*b} = [s^{*1}, s^{*2}, \ldots s^{*4600}]^b$ in the bth bootstrap sample. Thus four K-S D statistics are computed, one from each of the four fitted candidate models that are fit to data from the bth bootstrap sample. Of the four values of D, the minimum value, D_{Min}^{*b} is used as a measure to select which of the four candidate distributions represents the best fit to the statistics in that particular bootstrap sample. Thus the candidate distribution with the minimum D_{Min}^{*b} will be the candidate distribution for the bth bootstrap sample. For each of the B = 200 bootstrap samples, one well-known distribution will outperform the other three distributions to form a minimum D statistic vector, $D_{Min}^{*} = (D_{Min}^{*1}, D_{Min}^{*2}, \ldots, D_{Min}^{*b}, \ldots, D_{Min}^{*200})$, with 200 minimum Ds as its elements. The final chosen null distribution class will be the well-known candidate distribution that produces the largest number of minimum Ds in the vector D_{Min}^{*}. How this works will be seen in a later subjection that illustrates its use on data.

The minimum AIC and BIC criterion are also utilized to get the null distribution class which uses similar steps but changes the K-S statistics D to AIC and BIC. This criterion is used to obtain information in addition to the minimum D criterion for the purpose of selecting the best candidate models.

4.2.2. Determine the Final Parametric Null Distributions and Assess the Goodness-of-fit

In what follows, the bootstrap null distribution class is determined by the minimum K-S D criterion. Now the question is: How to determine the parameters for the chosen class distribution?

The final null distribution model will be represented by using three different parameter estimates. The first set of estimates is from the 5th percentiles of the 200 MLEs (i.e. 5th percentile parametric distribution), the second set is from the mean of the 200 MLEs (i.e. mean parametric distribution), and the last set is from the 95th percentiles of the MLEs (i.e. mean parametric distribution).

After the parametric null distribution is determined, its fit to the statistics, tg*, SSD*, and R_T^*, is assessed using empirical distributions obtained from the 200 bootstrap samples.

Figure 4.3 shows the work flow for determining the empirical distribution of the three test statistics from the bootstrap. Using the tg statistic in the bth bootstrap sample $\text{tg}^{*b} = [tg^{*1}, tg^{*2}, ..., tg^{*4600}]^b$ as an example:

1) Table (a) in Figure 4.3 shows the empirical distribution of the tg* statistic from the 200 bootstrap samples for 100 different percentiles (i.e., the 1st to the 100th). The 100 percentiles of the vector $\text{tg}^{*b} = [tg^{*1}, tg^{*2}, ..., tg^{*4600}]^b$ are calculated from the data in each row.
2) Suppose that 200 empirical cumulative density functions can be generated from each row in Table (a) of Figure 4.3. We need to use one empirical CDF to represent the empirical distribution for the statistic that is summarized from the 200 ECDFs. This final empirical CDF can be represented by three curves which are the 5th percentiles, median and 95th percentiles empirical distributions. The empirical 5th percentile empirical distribution is generated from the 5th percentiles of each column in Table (a) of Figure 4.3. Similarly,

the empirical median distribution is from the medians of each column, and the empirical 95th percentile distribution is from the 95th percentiles of each column in Table (a).

3) The final parametric null distribution is evaluated by the final empirical distribution by matching the corresponding 5th, median and 95th percentiles of their empirical distributions to their corresponding parametric distributions to see how close they are.

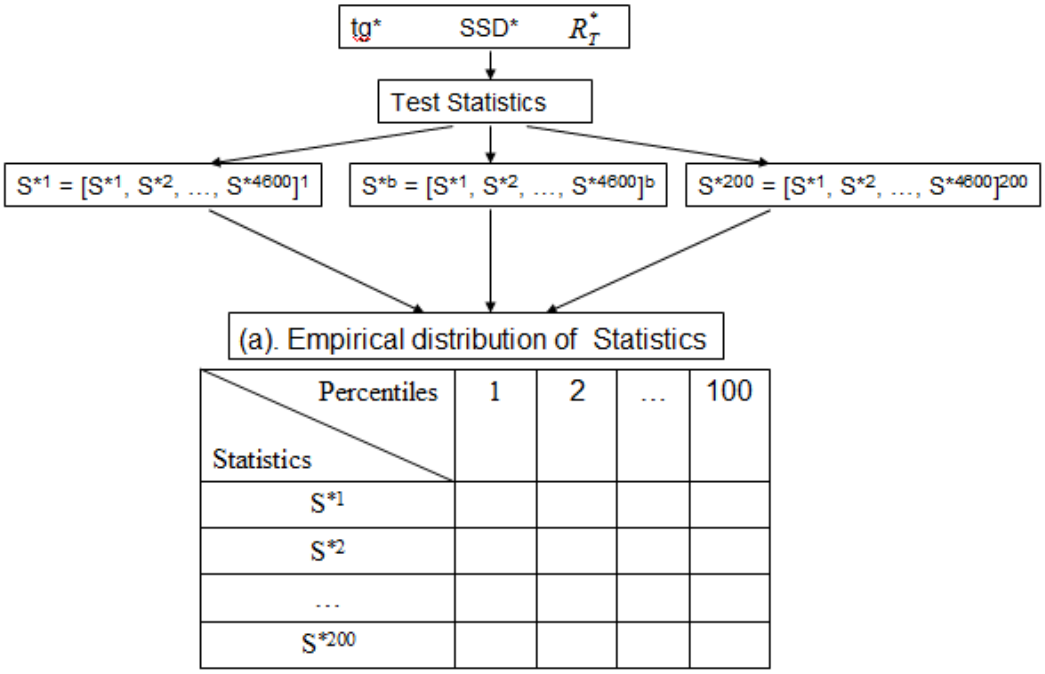

Figure 4.3: The empirical distributions of the three test statistics from 200 bootstraps

Using tg^* as an example, tg^{*b} is a vector with 4600 element, i.e., $tg^{*b} = [tg^{*1}, tg^{*2}, ..., tg^{*4600}]^b$ from the bth bootstrap. The empirical distribution of tg* will be listed in Table (a) using 100 percentiles (1 to 100). Similarly, the empirical distributions of the SSD* and R_T^* can be generated.

4.3. Results in *fad2* Dataset

4.3.1. Choose the Null Distributions

Table 4.1 shows the counts of minimum D statistics for the 4 candidate distributions from the 200 bootstrap samples in fad2 data. Among the 200 minimum D statistics in $D_{Min}^* = (D_{Min}^{*1}, D_{Min}^{*2}, ..., D_{Min}^{*b}, ..., D_{Min}^{*200})$, 101 minimum D statistics are from the exponential distribution. The exponential distribution is chosen to be the null distribution class for tg* since the exponential distribution outperforms the other three distributions by comparing 101 to 34, 48

and 17. By the minimum D criterion, the Weibull distribution is chosen for the statistics SSD^* and R_T^*.

Figure 4.4 shows the distribution of the real values of Kolmogorov-Smirnov test (K-S test) D statistics from each model fitting procedure. The left panel of Figure 4.4 shows the 4 K-S D statistics distribution for the statistic tg^*. Among the four D distributions, the exponential distribution contains smaller values of K-S D statistics compared to the other three well-known distributions. In the middle and right panels which plot the results of SSD^* and R_T^*, respectively, the Weibull distribution shows supporting evidence to be a better fit than other three well-known distributions with lower values of K-S D statistics.

Table 4.2 lists the final parametric null distributions for all the three test statistics tg, SSD and R_T in three forms: 5^{th} percentile of MLEs distributions, mean estimates distributions and 95^{th} percentiles of MLEs distributions. Note that in Table 4.2, the MLEs are the rate for the exponential distribution, shape and scale parameters for the Weibull distribution.

Table 4.1: The counts of the minimum K-S test statistic D_{Min} for the 4 candidate distributions.

In the first row for tg^*, among 200 minimum Ds in $D_{Min}^* = (D_{Min}^{*1}, D_{Min}^{*2}, ..., D_{Min}^{*b}, ..., D_{Min}^{*200})$, 101 minimum Ds are from the Exponential distribution, 34 minimum Ds are from the Gamma distribution, 48 minimum Ds are from the Lognormal distribution and 17 minimum Ds are from the Weibull distribution. The exponential distribution outperforms all other three distributions for tg^*. Hence, the exponential distribution is chosen to be the null distribution for tg^*.

	4 well-known distributions				
Statistics	**Exponential**	**Gamma**	**Lognormal**	**Weibull**	**Total**
tg^*	101	34	48	17	200
SSD^*	0	60	8	132	200
R_T^*	0	3	0	197	200

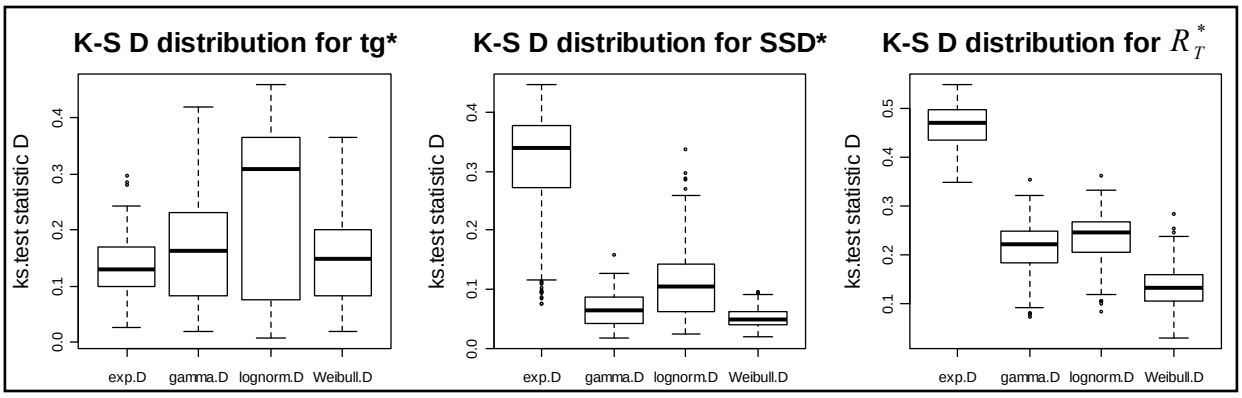

Figure 4.4: The distribution of K-S test statistic D for all three statistics tg*, SSD* and R_T^*
The box plots are from the values of the Kolmogorov-Smirnov test (K-S test) D statistics from the 200 bootstrap samples for each of the 4 well-known distributions. For example, in the left panel for statistic tg*, the 200 bootstrap D distribution is shown in the first box plot by fitting tg*s to the exponential distribution. The same tg* values are also fitted to the gamma, lognormal and Weibull distributions, and the distribution of their K-S test statistics D are shown with the other 3 box plots in the left panel, respectively.

Table 4.2: The final null distributions of the three statistics from the chosen parametric distribution with the parameter of estimates
The null distribution is shown in three forms, i.e., 5th percentile, mean and 95th percentile parametric distributions for each statistics. The final null distribution for the three statistics tg*, SSD* and R_T^* are the chosen distributions from the minimum K-S D criterion.

Statistics	5th percentile distribution	Mean distribution	95th percentile distribution
tg*	*Exp* (0.53)	*Exp* (0.73)	*Exp* (1.023)
SSD*	*Weibull* (1.41, 0.41)	*Weibull* (3.14, 0.82)	*Weibull* (5.81, 1.47)
R_T^*	*Weibull* (5.25, 3.19)	*Weibull* (8.68, 3.66)	*Weibull* (13.77, 4.50)

4.3.2. Assess the Final Selected Parametric Null Distributions using the Empirical Distributions and Fitting Results to the Data

Figure 4.5 shows how well the selected Weibull distribution fits the empirical bootstrap distributions for the statistic SSD in three forms. Its 5th percentile, median and 95th percentile empirical distributions fit to the 5th percentile, mean and 95th percentile the chosen Weibull distribution. Similar procedures can be applied to the statistics tg and R_T. Figure 4.5(a) shows that among the four well-known distributions the Weibull distribution has the smallest distance

to the median empirical CDF compared with the distances between all other curves to the median ECDF. The final parametric Weibull model appears to be better than the other parametric models. Figure 4.5(b) shows the distributions of 20 different percentiles (from the 100) using 20 box plots. The red curves are the 5th, median and 95th empirical null distributions which are from the 5th percentiles, medians and 95th percentiles of the 100 box plots. Figure 4.5(c) shows how well the final parametric Weibull distribution fits the empirical CDFs of the 20 box plots. The parametric Weibull distributions, including 5th, mean and 95th percentile distributions, are closely laid over the empirical box plots to their corresponding statistics. Figure 4.5(d) shows that the 3 parametric Weibull distributions can match the three empirical distributions very well.

The 95th percentile parametric null distributions are used to fit the three statistics in order to fit a mixture model. The 95th percentile distributions are used as a "bounding distribution" to represent a distribution that is analogous to the 95th percentile of a null distribution when testing a hypothesis with a univariate statistic. Characteristics of this bounding distribution are be explored as future work. More discussions of this can be found in chapter 10 topic.

Figure 4.6 shows the 95th percentile exponential distribution fitted to the statistic tg in fad2. The 95th percentile parametric exponential distribution *Exp* (1.023) captures well the shape of the tg data distribution in the right tail area. The 95th percentile parametric exponential distribution is the parametric fit to the pink histogram from the 95th percentile of the empirical distribution. The significant lipid pairs should come from those tg's that are close to 1, and that appears in the top area of the null distribution.

Figure 4.7 shows the 95th percentile *Weibull*(5.815, 0.472) distribution fitted to the statistic SSD in fad2. It is also from the parametric fit to the pink histogram from the 95th percentile of the empirical distribution. Since the larger values of SSD are of most interest for the lipid pairs that are significantly affected by the mutation, the area on the right of the 95th percentile Weibull distribution should include the significant results.

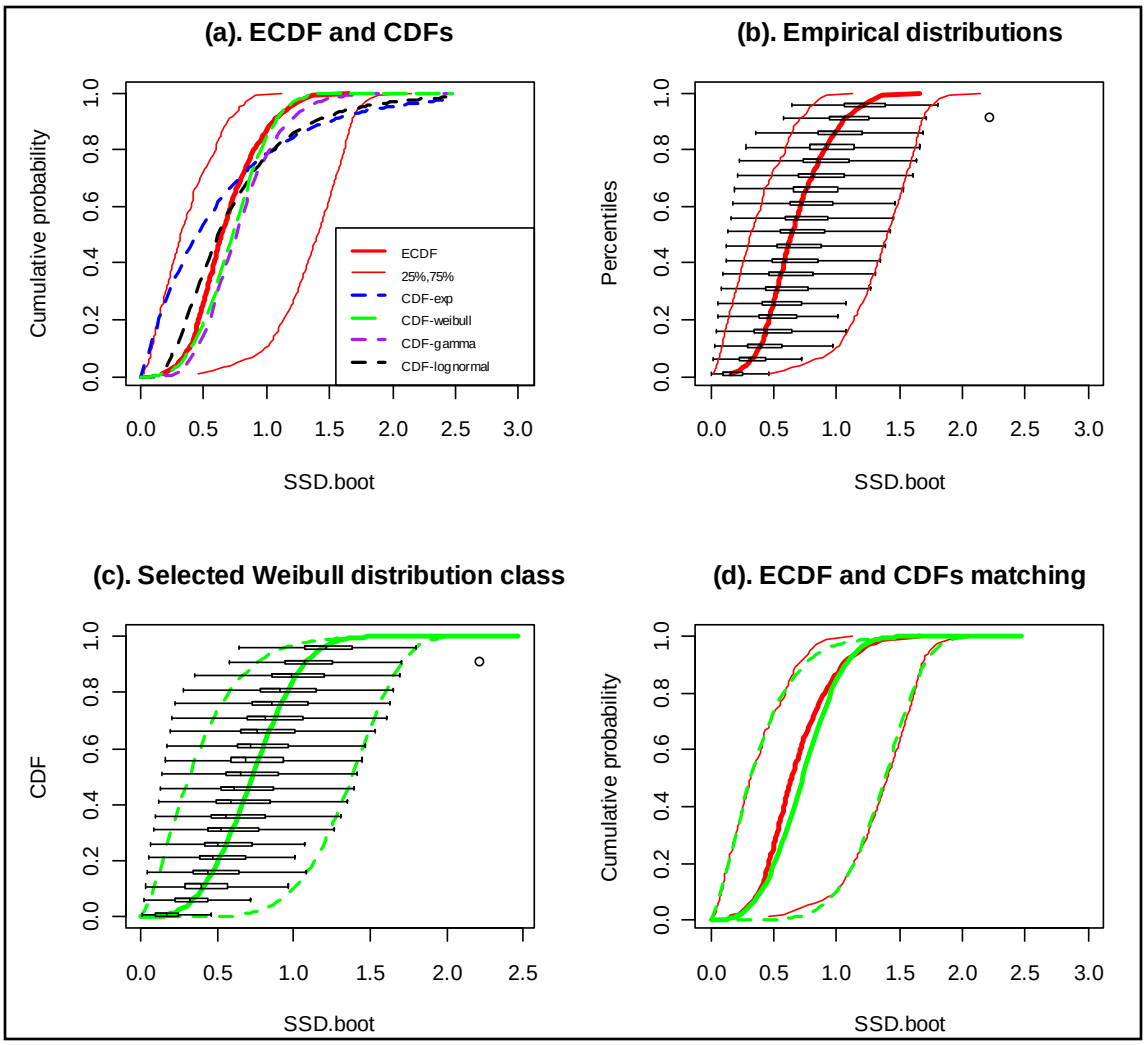

Figure 4.5: ECDFs matches with the selected null Weibull CDFs for the statistic SSD
(a): The thicker red line is the empirical CDF of the bootstrap null distribution which is summarized from the median distribution that is described in Figure 4.3 in determining the empirical distributions. The two thinner red lines are the 25th and 75th percentiles empirical distributions. The two thin red lines form a 50% region to capture the null distributions from the parametric null distributions that is shown in 4 dashed curves. The 4 dashed curves are exponential (blue), Weibull (green), gamma (purple) and lognormal (black) parametric null distribution, respectively. **(b):** The series of box plots show the empirical distributions from 20 percentiles out of the 100 percentiles in Table (a) of Figure 4.3. The three red curves represent the three empirical distributions, 5th, median and 95th for SSD for panel (a). **(c):** The same 20 box plots from panel (b) are matched with the chosen parametric null Weibull distribution in three forms: 5th, mean and 95th parametric null distributions. **(d):** The final parametric Weibull null distribution is matched with the final empirical null distribution in three forms. The three red curves are the empirical null distribution and the three dashed curves are the three parametric null distributions that are chosen by fitting the parametric model to the bootstrap statistics.

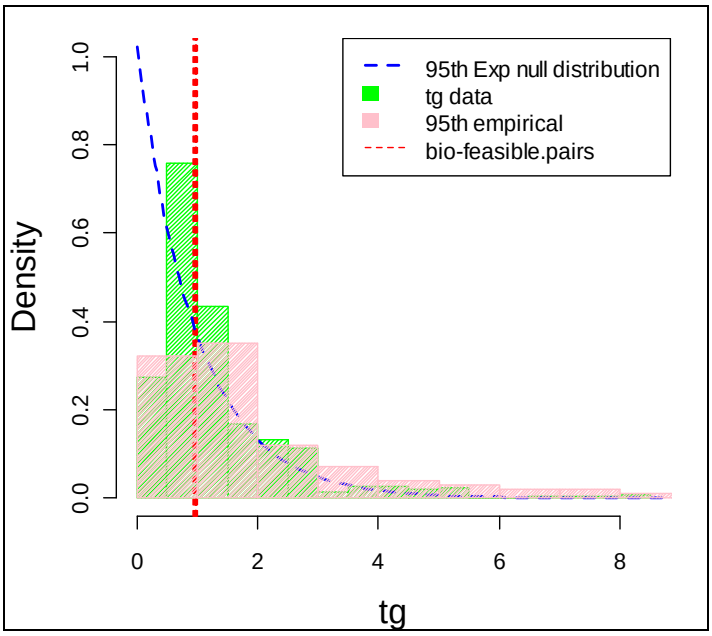

Figure 4.6: tg statistic parametric bootstrap null distribution overlaid with the tg distribution in fad2

The red dotted line shows the tg statistics of all biologically functional AB pairs in fad2 which are close to 1. The green histogram represents the tg distribution in fad2. The dashed green curve is the 95th parametric null $Exp(1.023)$ that fits to the pink 95th empirical histogram.

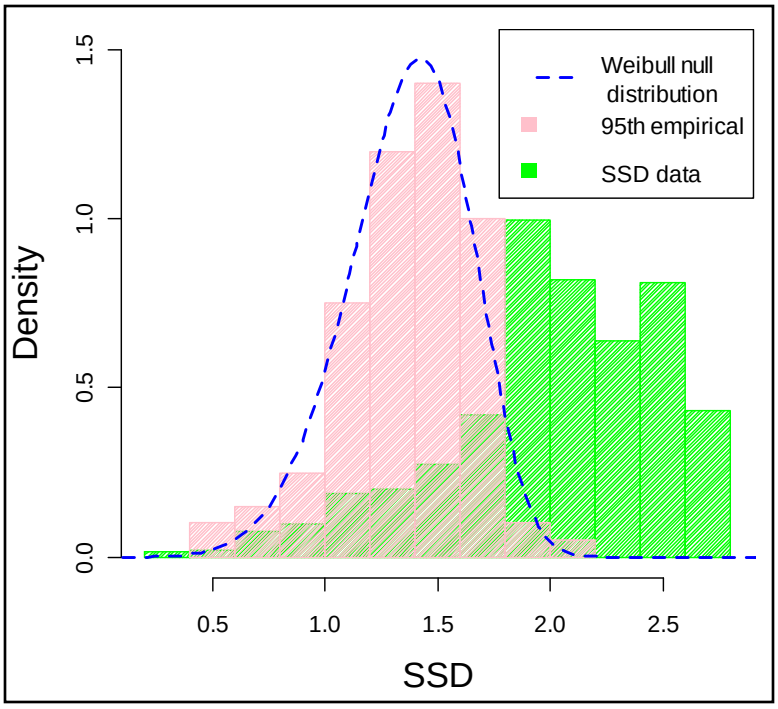

Figure 4.7: SSD parametric bootstrap null distribution overlaid with the SSD distribution in *fad2*

The green histogram is the SSD distribution in fad2. The pink histogram is the 95th percentiles of the empirical bootstrap data. The blue dashed line is the parametric 95th null $Weibull(5.81, 1.47)$ that fits to the pink 95th empirical histogram.

Figure 4.8 shows the overlaid plot of the selected 95th percentile $Weibull(13.77, 4.50)$ null distribution with the real data distribution for statistic R_T in fad2 data. The 95th percentile parametric distribution did not fit the pink histogram from the 95th percentile of the empirical distribution very well. Since R_T is a combined statistic using both tg and SSD, the bigger values of SSD may affect the R_T parametric fitting. Since larger values of R_T can give significant lipid pairs, the lipid pairs appearing on the right tail show the results of interest.

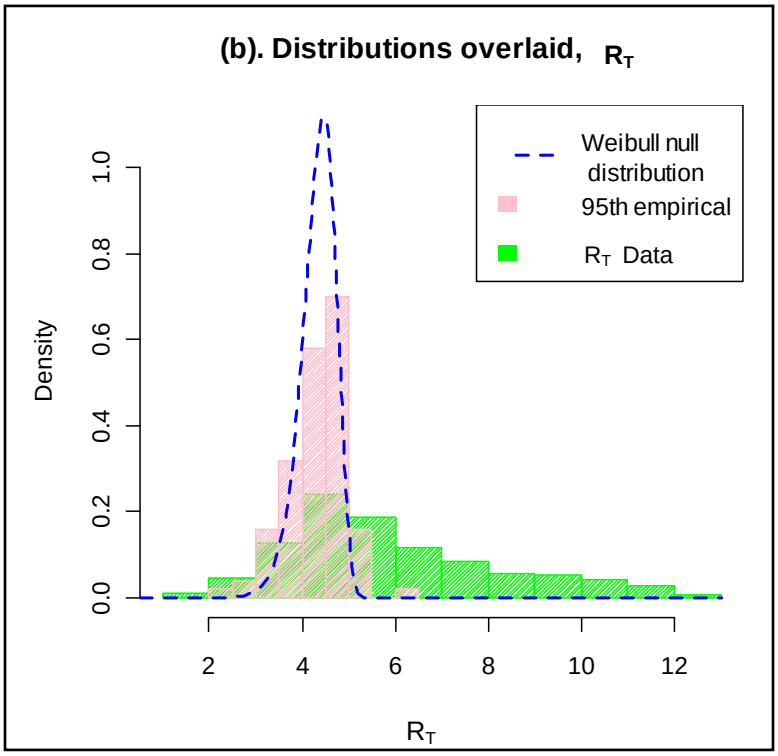

Figure 4.8: R_T bootstrap null distribution overlaid with the distribution in fad2
The green histogram is the R_T distribution in fad2. The dashed blue curve is the 95th parametric null distribution $Weibull(13.77, 4.50)$ that fits to the pink 95th empirical histogram.

Some summary remarks

A problem appeared in chapter 4 due to strong signal that is detected in the fad2 dataset. When the signal is strong (i.e. a strong mutation effect on lipid concentrations), a single parametric may be inadequate to capture the shape of an empirical distribution. Another technique will be explored in a later chapter. The bootstrap procedure is applied under a restrictive null hypothesis of $F = G$. Thus any treatment affecting most of the lipidome is likely to depart significantly from the null hypothesis under which the bootstrap procedure was carried

out. Still, it is worth noting that the techniques described herein should be appropriate in cases where the real interest in testing the high-dimensional null hypothesis given in (4.1).

Chapter 5 - A Mixture Normal Bootstrap Null Distribution

Chapter 4 investigated the parametric bootstrap null distribution (PBN) fitting to the bootstrap test statistics tg, SSD and R_T. There was a strong signal in fad2 data. To assess the mutation effects in the lipid pairs, it is important to know how strong the treatment effect is when comparing the MT treatment effect with the WT (control) in each dataset. The choice of technique for modeling a null distribution may change depending on the overall effect of the mutation on the entire lipidome. To find a proper null distribution to fit the empirical bootstrap distribution of the test statistic, a mixture normal bootstrap null distribution (MNBN) will be investigated in this chapter.

5.1. Introduction to the Mixture Normal Distribution

The history of mixture normal distributions can be dated back to the nineteenth century by Simon Newcomb and Karl Pearson (Wirjanto and Xu, 2009). Since then, mixture normal models have been widely used in biology, engineering, economics and other applied areas. Applications of the mixture normal distributions can be found in Everitt and Hand (1981), Titterington et.al. (1985), and Mclachlan and Peel (2000). The focus of the mixture normal distribution has been on the parameter estimation, and testing the number of components in the mixture normal models. Some R packages deal with finite mixture model fitting problems. In the R package "mixtools" (*http://cran.r-project.org/web/packages/mixtools/mixtools.pdf*), the authors developed a set of R objects to find the Maximum Likelihood Estimates of the parameters for a finite mixture model. One advantage of the mixture normal distribution is its great flexibility in capturing shapes. The mixture normal distributions can capture multimodal or skewed continuous distribution very well. In this chapter, the flexibility of the mixture normal is utilized in capturing the shape of the bootstrap null distribution.

5.2. Mixture Normal Bootstrap Null Distribution (MNBN)

The focus here will be on the bootstrap null distribution of R_T^* that was produced under the null hypothesis H_0: F = G, because this test statistic includes information in both metrics, tg* and SSD*. Let $x = R_T^*$. The probability density function for a normal distribution is

$$f(x/\mu,\sigma)=\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{5.1}$$

where $-\infty < x < +\infty$, the mean parameter $-\infty < \mu < +\infty$ and standard deviation $\sigma > 0$. The probability density function of the mixture normal distribution has the form of (2.5) with k normal distributions as its components. The mixture model will be used to fit the data using maximum likelihood estimation (MLE). So, technically, we are assuming that roughly 4,600 lipid pairs satisfying the screening for $y^* = 2$ are mutually independent. This assumption is likely not met in a lipidomics experiment. Nevertheless, it has been made in many high-dimensional "omics" studies for purposes of fitting a mixture distribution (for more references, see Gadbury et al., 2008).

Let i be the number of lipid pairs in the dataset, and $i = 1, 2, \ldots, M$, where M is around 4600, depending on the dataset. The research interest is that H_{0i}: The i^{th} reactant and product pair is not affected by the mutation, i.e. F = G versus H_{ai}: The i^{th} reactant and product pair is affected by the mutation, i.e., F ≠ G. The likelihood function for the R_T^* statistic from a normal mixture model with k components under the null hypothesis can be expressed as

$$L=\prod_{i=1}^{M}\left[\sum_{j=1}^{k}\pi_j\cdot f_j(x_i/\mu_j,\sigma_j)\right], \tag{5.2}$$

where x_i is the R_T^* statistic for the i^{th} test. π_j is the mixing proportions on each of the component densities, satisfying the constraints $\sum_{j=1}^{k}\pi_j = 1$, and $0 \le \pi_j \le 1$.

5.3. Randomization Test for the Treatment Effects

The randomization test is utilized to test the strength of the mutation effects in all 9 experiments. The results from six fad datasets fad2, fad3, fad4, fad5, fad6 and fad7 are used as an illustration. In each dataset 5 samples are from the WT group and 5 samples are from the MT group. $\binom{10}{5} = 252$ distinct treatment arrangements are chosen to calculate the randomization p value for each lipid. The proportion of the number of smallest p values out of the total number of p values in each dataset is used to measure the strength of the mutation effect on the FAD genes. Table 5.1 shows the number of smallest p values, the total number of p values, and the resulting

proportion in each dataset. Let $\bar{x}_{wt}$ and $\bar{y}_{mt}$ denote the sample means for one lipid in the WT and MT groups. Let $\bar{x}^*_{wt}$ and $\bar{y}^*_{mt}$ denote the group means for one lipid in the permutation samples. The p values from the two-tailed randomization test are calculated with $\frac{\#(|d^*| > |d|)}{252}$, where $d^* = \bar{x}^*_{wt} - \bar{y}^*_{mt}$ is the mean difference between the WT and MT groups in one lipid for the permutation samples, and $d = \bar{x}_{wt} - \bar{y}_{mt}$ is the observed mean difference between the WT and MT groups for the same lipid. The smallest p value is $\frac{1}{126} = 0.0079$.

Table 5.1: Number of the smallest p values, number of p values, and proportion of the smallest p values in each dataset

	fad2	fad3	fad4	fad5	fad6	fad7
# of smallest p values	76	47	10	28	43	35
# p values	136	127	129	130	135	132
Proportion	0.56	0.37	0.08	0.22	0.32	0.27

From Table 5.1, we can see that fad4 has the lowest proportion of the smallest p values and fad2 has the largest proportion of the smallest p values. Hence, fad4 dataset is expected to have the smallest overall mutation effect on the lipid concentrations, while *fad2* has the highest mutation effect. Figure 5.1 shows the p value distribution from the randomization test for each dataset using bar charts. If there is strong mutation effect, we expect that the p value distribution will be a right-skewed distribution with a peak on the smallest p value, 0.0079. Otherwise, the p value distribution should be close to a flat shape, showing a weaker treatment effect. From Table 5.1 and Figure 5.1 we can see that the strength of the signal from the strongest to the weakest from all the datasets should follow the order: fad2, fad3, fad6, fad7, fad5, and fad4. In the following parts of this thesis, the datasets fad2 and fad4 would be used as an illustration for the mixture model to fit the bootstrap null distribution.

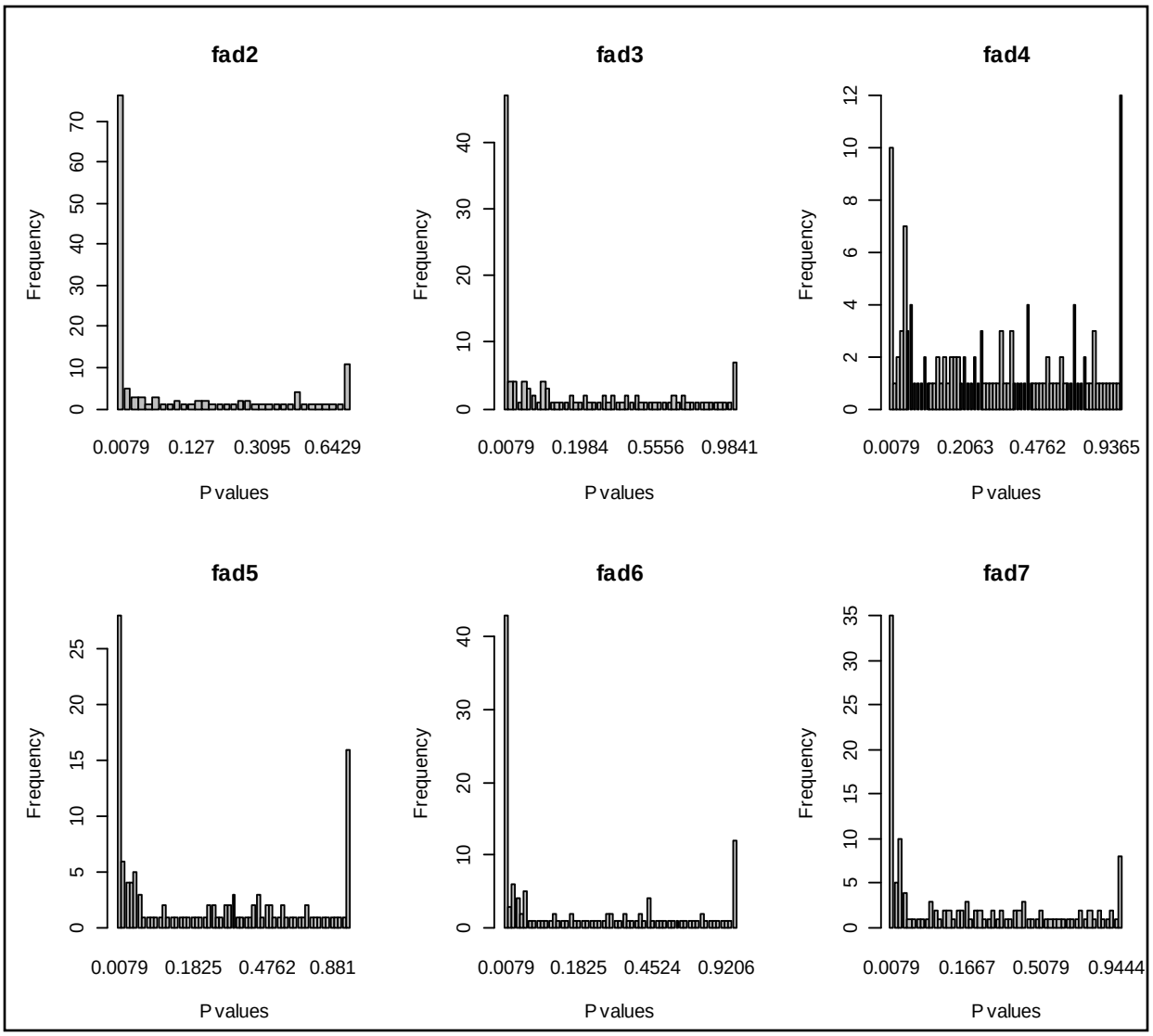

Figure 5.1: Bar charts for the distribution of the p values from the randomization test
The first bar on the left in each plot shows the frequency of the smallest p value 0.0079. The last bar shows the frequency of the biggest p value 1 in the randomization test.

5.4. MNBN in fad2 and fad4 Datasets

5.4.1. Find the MNBN Distributions

To investigate how well the mixture normal model fits the bootstrap null distribution, the mixture normal model in (5.2) is used to fit the bootstrap null distribution of R_T from Chapter 4 which was produced from the 200 bootstrap samples under the null hypothesis H_0: F = G. Since fad2 and fad4 are the datasets who show extremely big or small mutation

effects, datasets fad2 and fad4 will be used in the illustration. Figure 5.2 shows the bootstrap null distribution overlaid with the R_T distribution in fad 4.

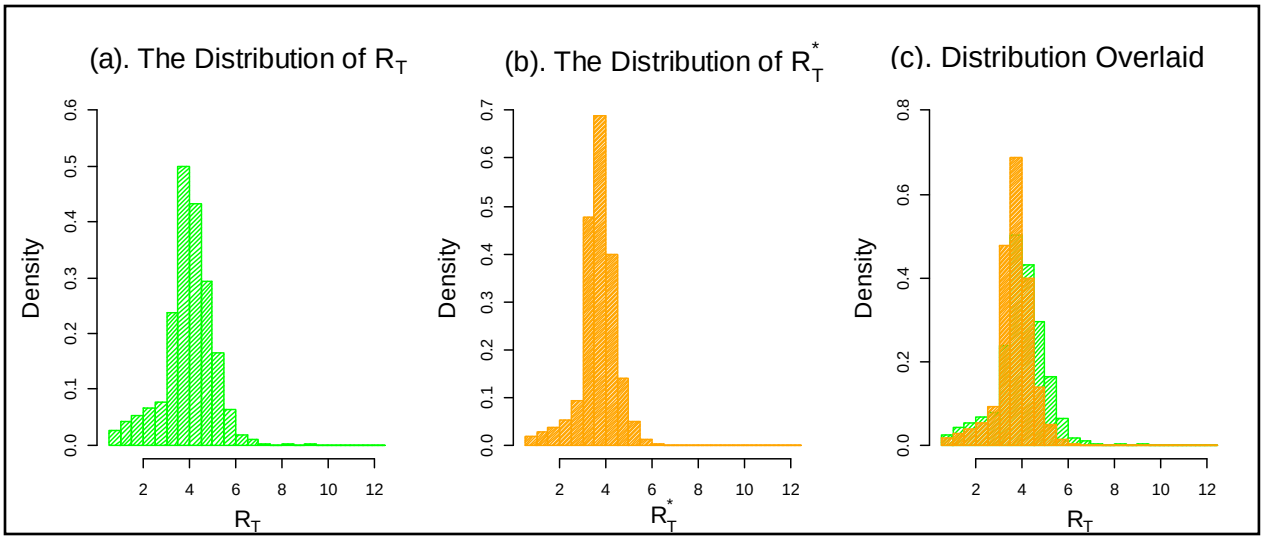

Figure 5.2: The bootstrap distribution overlaid with the R_T distribution in fad4 data
(a). The R_T statistics distribution from the fad4 data. (b). The empirical bootstrap null distribution which contains 4600×200 R_T^* statistics. Each bootstrap sample contains about 4600 R_T^* statistics. (c). The bootstrap distribution, R_T^*, (orange) overlaid with the data, R_T, (green).

Figure 5.2(a) shows the R_T distribution in fad4, and Figure 5.2(b) is the histogram including all the bootstrap samples from the 4600×200 R_T statistics. The big vector of R_T statistics is overlaid in Figure 5.2(c). From Figure 5.3(c), we can see that the actual bootstrap null histogram covers a large portion of the data histogram. In chapter 4, a parametric distribution, which was generated by averaging the parameters obtained from fit of all the bootstrap samples, was used to fit the actual bootstrap sample. However, the selected parametric distribution in this way may miss the actual shape of the null distribution. Therefore, an alternative approach using a normal mixture model may better capture the shape of the bootstrap null distribution.

Next, we need to focus on how to fit the mixture normal model to the orange histogram in figure 5.2 for getting a null distribution Parameter estimates for the five parameters π_1, μ_1, σ_1, μ_2, σ_2 are shown in Table 5.2. Figure 5.3 shows the two component mixture normal distribution fit to the bootstrap samples in fad4. Let $x = R_T^*$. The MNBN distribution with two components in fad4 can be written as

$$f_{H_0}(x) = 0.28 \cdot N(3.44,1.26) + 0.72 \cdot N(3.76,0.44).$$

Table 5.3 shows the MLEs of the estimates for a three component mixture normal null distribution in fad4. Figure 5.4 is the graphical illustration of the three component mixture normal distribution from Table 5.3.

Table 5.2: The MLEs and the log-likelihood value for mixture normal null distribution in fad4 with two components and five parameters $\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2$.

Parameter π_1 is the proportion weight for the first component. μ_1 and μ_2 are the means for the two components. σ_1 and σ_2 are the standard deviations for both components.

Parameters	Estimates	Standard Error	95% confidence interval	Log-likelihood
π_1	0.28	0.0071	(0.269, 0.297)	-35325.43
μ_1	3.44	0.0132	(3.413, 3.465)	
μ_2	3.76	0.0038	(3.755, 3.77)	
σ_1	1.26	0.0138	(1.232, 1.286)	
σ_2	0.44	0.0036	(0.432, 0.446)	

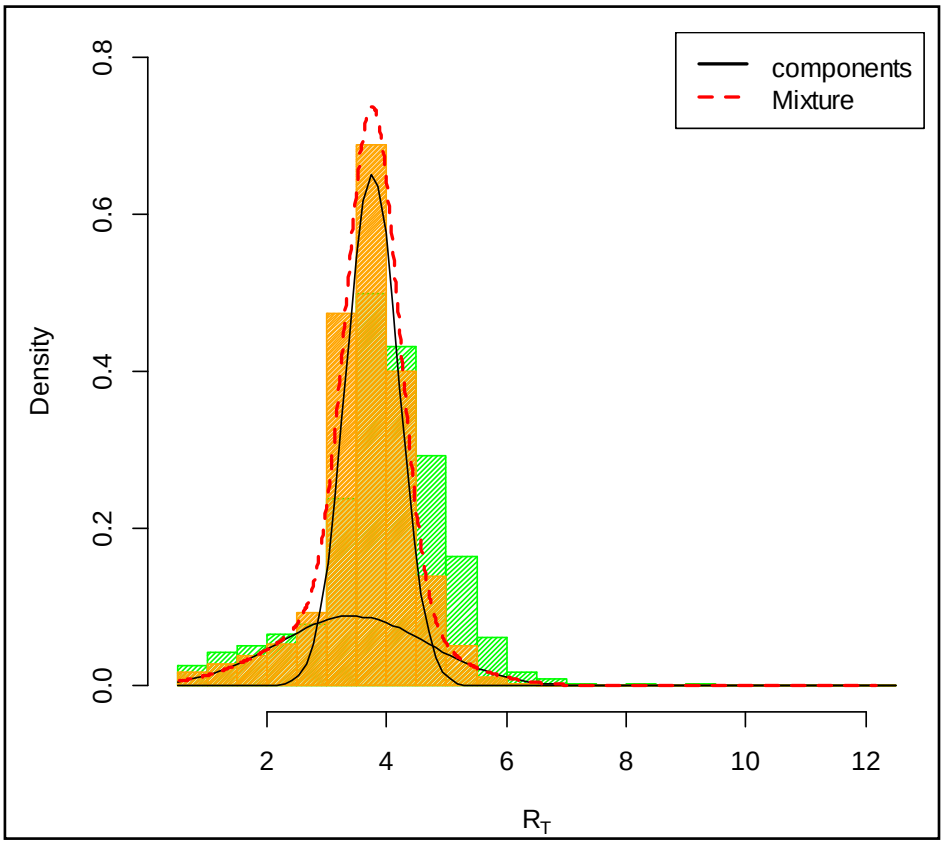

Figure 5.3: Two-component MNBN distribution of R_T^* in fad4 data

The orange histogram is the bootstrap null distribution which contains about 4600×200 R_T^* statistics. The two black solid curves are the two normal components. The red dashed curve is the mixture model from the two normal components.

Table 5.3: The MLEs and the log-likelihood value for mixture normal null distribution in fad4 with three components and seven parameters $\pi_1, \pi_2, \mu_1, \sigma_1, \mu_2, \sigma_2, \mu_3, \sigma_3$.

Parameter π_1 and π_2 are the proportion weights for the first two components. μ_1, μ_2 and μ_2 are the means for the three components. σ_1, σ_2 and σ_3 are the standard deviations for the three components.

Parameters	Estimates	Standard Error	95% CI	Log-likelihood
π_1	0.603	0.030	(0.543, 0.662)	
π_2	0.190	0.029	(0.132, 0.247)	
π_3	0.208	0.007	(0.195, 0.221)	
μ_1	3.670	0.008	(3.654, 3.686)	
μ_2	4.287	0.081	(4.127, 4.446)	-35050.9
μ_3	3.112	0.029	(3.054, 3.170)	
σ_1	0.380	0.006	(0.369, 0.391)	
σ_2	0.552	0.032	(0.489, 0.616)	
σ_3	1.252	0.015	(1.223, 1.281)	

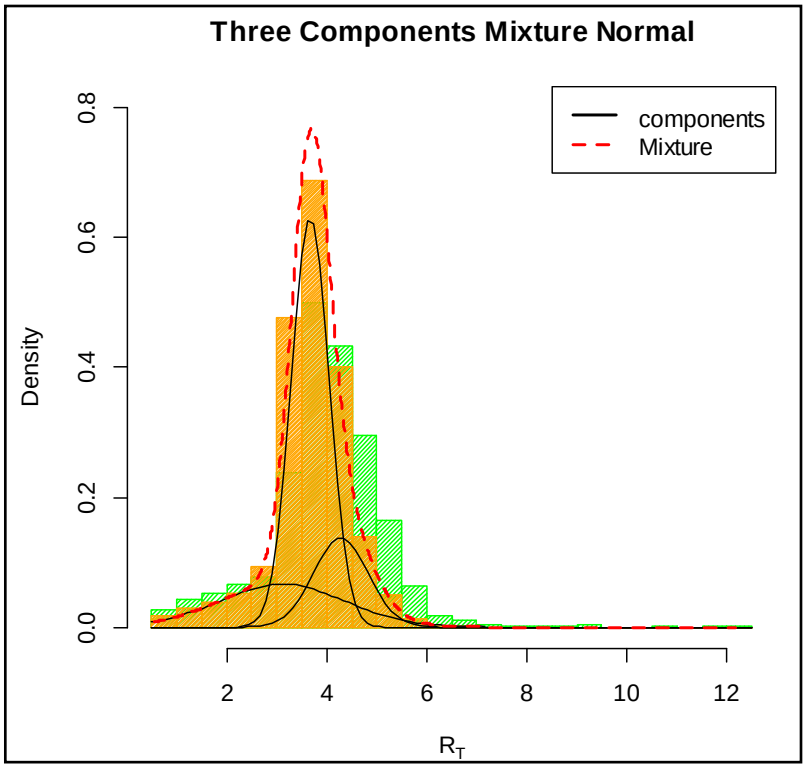

Figure 5.4: Three-component MNBN distribution of R_T^* in fad4 data

The three black solid curves are the three normal components. The red dashed curve is the mixture model from the three normal components.

The MNBN distribution with three components in fad4 can be written as

$$f_{H_0}(x) = 0.602 \cdot N(3.67, 0.38) + 0.208 \cdot N(3.11, 1.25) + 0.19 \cdot N(4.29, 0.55).$$

From the two component mixture normal distribution shown in Table 5.2, we can see that the 95% confidence intervals (3.413, 3.465) and (3.755, 3.77) do not overlap. This provides one indication that two components are needed. However, in Table 5.3 the 95% confidence intervals for the three means (3.654, 3.686), (4.127, 4.446), and (3.054, 3.170) do not overlap, either. The question is how to choose the number of components? Allison et al. (2002) suggested a bootstrap method to test the number of components in the mixture model using a statistic $Q = 2(L_\nu - L_{\nu-1})$, where L_ν and $L_{\nu-1}$ are the log-likelihood functions of the mixture model with ν and $\nu - 1$ components, respectively. Another bootstrap method that is analogous to the method used in Allison et al. (2002) will be proposed in chapter 7 to test the number of components in a normal mixture model. Here, two component MNBN distribution will be used for further analysis since the two-component mixture model can capture the shape of the empirical bootstrap null distribution.

5.4.2. The Results from fad4 and fad2

The hypotheses testing would be conducted for H_0: F = G, if there is no significant mutation effect in each lipid pair, versus Ha: $F \neq G$ assuming there is significant mutation effect in each lipid pair. The two-component $f_{H_0}(x) = 0.28 \cdot N(3.44, 1.26) + 0.72 \cdot N(3.76, 0.44)$ MNBN distribution in fad4 is utilized to find the p values for testing if the lipid pair is effected by the mutation. Since bigger R_T statistics show evidence for a real finding, an upper-tail test will be appropriate to get a list of significant findings. The p values are calculated as

$$P(X > x_{observed}) = \int_x^{+\infty} f_{H_0}(x)dx = \int_x^{+\infty} \left[0.28 \cdot N(3.44, 1.26) + 0.72 \cdot N(3.76, 0.44)\right]dx.$$

where $x_{observed}$ = R_T are the observed statistics in the data. The local *fdr* multiple adjusting procedure is applied to control the false discovery rate in a family of 3000 lipid pairs in fad4.

Table 5.4 shows a list of significant findings after the local *fdr* adjustment. Other multiple adjustment procedures holm, hochberg, bonferroni, BH and BY (Holm, 1979; Hochberg, 1988; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) are also explored. The method BY which was proposed by Benjamini and Yekutieli (2001) controls the false discovery rate in multiple testing under dependency.

Table 5.4: The list of significant findings using the MNBN distribution under the null hypothesis F = G

The red lipid pairs are the biologically functional lipid pairs in fad4. The R_T is the test statistic for that lipid pair. Pval is the p values. The rest of the columns are the adjusted p values using fdr, holm, hochberg, bonferroni, BH and BY multiple testing adjustment procedures.

Lipid Pairs	R_T	Pval	fdr	holm	hochberg	bonferroni	BH	BY
PG34_3_PG34_4	12.182	5.43E-13	1.49E-08	1.61E-09	1.61E-09	1.61E-09	1.61E-09	1.38E-08
PG34_3_lysoPG18_3	11.702	7.50E-12	9.99E-07	2.22E-08	2.22E-08	2.22E-08	1.11E-08	9.53E-08
PG34_3_PG32_1	10.885	4.75E-10	0.000314	1.41E-06	1.41E-06	1.41E-06	4.70E-07	4.03E-06
PG32_0_PG34_4	9.412	2.98E-07	0.0003	0.0009	0.0009	0.0009	0.0002	0.0019
PG32_0_lysoPG18_3	9.341	3.92E-07	0.0003	0.0012	0.0012	0.0012	0.0002	0.0019
PG34_3_PG34_2	9.310	4.42E-07	0.0007	0.0013	0.0013	0.0013	0.0002	0.0019
PG32_0_PG32_1	9.170	7.55E-07	0.0143	0.0022	0.0022	0.0022	0.0003	0.0027
PG32_0_PG34_2	8.530	7.46E-06	0.0198	0.0221	0.0221	0.0221	0.0028	0.0237
PG34_0_PG34_4	8.147	2.61E-05	0.0198	0.0773	0.0773	0.0775	0.0084	0.0724
PG34_0_lysoPG18_3	8.120	2.85E-05	0.0198	0.0842	0.0842	0.0845	0.0084	0.0724

Table 5.4 shows that there are 10 significant lipid pairs in fad4 by using local *fdr* adjustment. Note that there will be less significant results if we use other multiple adjustment procedures. The pairs which are highlighted in red are the biologically functional lipid pairs provided by the biologists in the Lipidomics Research Center at Kansas State University. In total there are 7 biologically functional lipid pairs, and the MNBN method can capture 6 of them. This method may accurately catches the significant findings compared to the method in Chapter 4 using the PBN which has 551 significant lipid pairs. The same MNBN procedure is applied to fad2 data. The significant findings include 2626 lipid pairs. The results from fad2 is 3207 lipid pairs using the PBN method.

In Chapter 6, we would check whether we can improve the method by using the bootstrap method under a different assumption, $\mu_F = \mu_G$, which assumes that the population means are the same for both the wild type group and the mutant group data.

Some summary remarks

In this chapter, we have used the Mixture Normal Bootstrap Null (MNBN) distributions to find the list of significant lipid pairs. MNBN is used to fit the empirical bootstrap data under the null distribution when F = G. We show that the MNBN distribution performs better than the PBN distribution used in Chapter 4 because the former method can more precisely select the most significant lipid pairs. We also show that the MNBN method seems preferred especially for the datasets whose mutation effects are weak, like fad4. For strong mutation effect datasets like fad2, this method can also improve the results to some extent. This is because the normal mixture model has more flexibility to capture usual shape and thus can be more discriminating in identifying real results in a large list of findings.

Chapter 6 - Bootstrap Methods Under the Equal Mean Hypothesis

In chapter 4 a parametric method was used to find the parametric bootstrap null (PBN) distribution from the bootstrap sample under a restrictive null hypothesis H_0: F = G, which assumes that the underlying distributions are the same in the wild type and mutant groups if there is no mutation effect. Under the same null hypothesis, the mixture normal bootstrap null distribution (MNBN) was investigated in chapter 5. The MNBN method provides more accurate results than PBN, especially in the datasets where the mutation effects are weak. Efron and Tibshirani (1993) illustrated a bootstrap method for testing equality of means of two treatment groups using a test statistic. This can be an alternative way to generate the bootstrap samples.

In this chapter the bootstrap samples would be generated under the equality of the means $\mu_F = \mu_G$, where μ_F and μ_G are the population means of the WT and MT groups in a lipid pair, respectively. The test statistics R_T^* will be generated in a similar way as in chapter 4. Both PBN and the MNBN methods will be investigated under the null hypothesis $H_0 : \mu_F = \mu_G$.

6.1. Bootstrap Algorithm Under Equal Mean Hypothesis

The Bootstrap procedure to get the empirical null distribution under the assumption $\mu_F = \mu_G$ is given below.

1. Let the sample be a 141 by 10 matrix with the 141 rows representing the 141 lipid species and the 10 columns representing 5 WT samples and 5 MT samples.
2. Let WT sample data be $w_1, w_2, ..., w_n$ with population mean μ_F and sample mean $\overline{w}$. Let the MT data be $m_1, m_2, ..., m_n$ with population mean μ_G and sample mean $\overline{m}$.
3. Transform the data by using $\widetilde{w}_i = w_i - \overline{w} + \overline{x}$ and $\widetilde{m}_i = m_i - \overline{m} + \overline{x}$, $i = 1,2,...,n$. Where $\overline{x}$ is the mean of the combined sample.
4. Let B = 200, and $b = 1, 2, ..., B$ to denote the b[th] bootstrap sample ($\widetilde{w}^{*b}, \widetilde{m}^{*b}$). Where $\widetilde{w}^{*b}$ is sampled with replacement from $\widetilde{w}_1, ..., \widetilde{w}_5$ and $\widetilde{m}^{*b}$ is sampled with replacement from $\widetilde{m}_1, ..., \widetilde{m}_5$.

5. Center and scale samples $\mathrm{x}^{*b} = (\tilde{w}^{*b}, \tilde{m}^{*b})$ by using $z_{ij}^{*} = \frac{x_{ij}^{*} - \bar{x}_{\bullet\bullet}^{*}}{s^{*}}$ to get the mean 0 and standard deviation 1 for each lipid species. x_{ij}^{*} and z_{ij}^{*} denote the bootstrap data before and after scaling.
6. Pair any reactant A^{*} with any product B^{*} to get approximately $2\binom{141}{2} = 19740$ lipid pairs $\mathrm{A}^{*} \rightarrow \mathrm{B}^{*}$.
7. Screen the relationships of $\bar{z}_{A1\bullet}^{*} < \bar{z}_{A2\bullet}^{*}$ and $\bar{z}_{B1\bullet}^{*} > \bar{z}_{B2\bullet}^{*}$ according to the value $y^{*} = 2$ for any arbitrary lipid pairs. In total, K = 4600 lipid pairs satisfy the conditions of $y^{*} = 2$.
8. Calculate the statistic $R_T^{*} = -\log(R^{*})$ for each lipid pair $\mathrm{A}^{*}\mathrm{B}^{*}$ from the scaled bootstrap sample z^{*b} with y^{*} = 2. The test statistic is a vector with about 4600 elements from each of the b^{th} bootstrap samples, i.e. $\mathrm{R}_{\mathrm{T}}^{*\mathrm{b}} = (\mathrm{R}_{\mathrm{T}}^{*1}, \mathrm{R}_{\mathrm{T}}^{*2}, ..., \mathrm{R}_{\mathrm{T}}^{*4600})$. The R^{*} statistic is defined as
$$\mathrm{R}^{*} = (tg^{*} - 1)^2 + (SSD^{*} - 2.684))^2 .$$
9. Use the 4600 bootstrap statistics to determine the null distribution for the statistic R_T^{*}.

Note that in step 3, the means of the WT and MT groups are first centered at zero by subtracting the corresponding group means. Then, the overall mean of the each lipid is shifted to $\bar{x}$ by adding the overall mean $\bar{x}$ to each of the ten centered WT and MT data. Also, centering and scaling the data do not change the correlation structure.

Again datasets fad2 and fad4 are used here. The former shows the strongest mutation effect, while the latter the weakest, as shown in the randomization test in chapter 5. In the following sections, these two datasets will be used as examples to illustrate both the PBN and MNBN methods under the null hypothesis $H_0 : \mu_F = \mu_G$. Before further analyzing the two datasets, it will be interesting to check whether the R_T^{*} distributions are the same in the bootstrap samples under the two different assumptions of $F = G$ and $\mu_F = \mu_G$. Figure 6.1 shows the comparison of the R_T^{*} bootstrap distributions of fad2 and fad4 under the two above-mentioned null hypotheses. Table 6.1 shows the means and the five number summary of the R_T^{*} bootstrap distributions under the two hypotheses for fad2 and fad4 datasets.

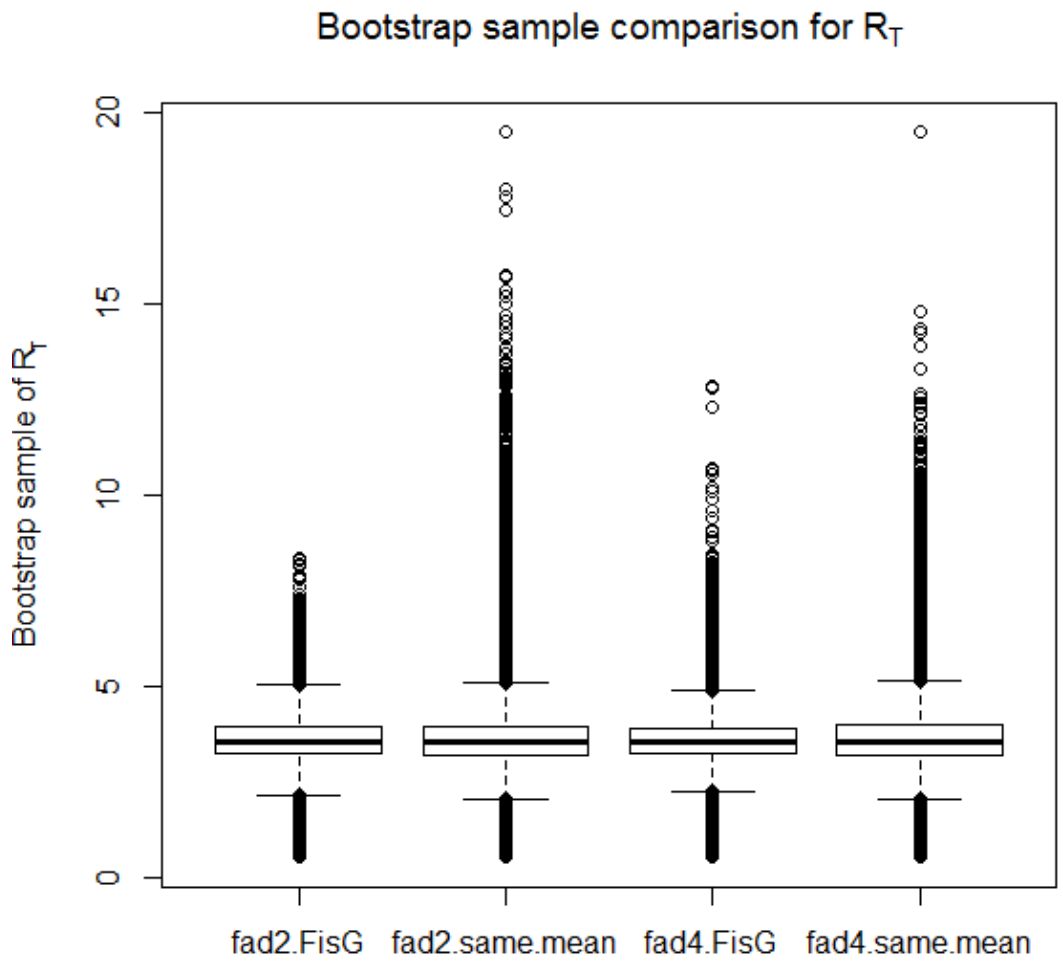

Figure 6.1: The R_T^* bootstrap distribution comparison under the assumptions of $F = G$ and $\mu_F = \mu_G$

The two box plots on the left are the R_T^* bootstrap distributions from fad2 under the assumptions $F = G$ and $\mu_F = \mu_G$, respectively. The two box plots on the right are the R_T^* bootstrap distributions from fad4 under the assumptions $F = G$ and $\mu_F = \mu_G$, respectively.

Table 6.1: The means and the five number summary of the R_T^* bootstrap sample distributions in fad2 and fad4 datasets under the two assumptions $F = G$ and $\mu_F = \mu_G$.

		Min	1st Qu.	Median	Mean	3rd Qu.	Max.
fad2	$H_0 : F = G$	0.52	3.23	3.54	3.55	3.96	8.35
	$H_0 : \mu_F = \mu_G$	0.52	3.20	3.53	3.55	3.96	19.48
fad4	$H_0 : F = G$	0.52	3.23	3.53	3.49	3.90	12.86
	$H_0 : \mu_F = \mu_G$	0.53	3.21	3.54	3.58	3.99	19.48

From Figure 6.1 and Table 6.1 we can see that the distributions of R_T^* in fad2 and fad4 are similar. The R_T^* distributions under $\mu_F = \mu_G$ are heavier upertailed than those under $F = G$ in both datasets fad2 and fad4. The third quartiles and the maxima are bigger under $H_0 : \mu_F = \mu_G$. Since there are more larger values of R_T^* statistics in the null distribution under $H_0 : \mu_F = \mu_G$, we investigate whether the null distribution under this hypothesis improves the results by giving a shorter list of significant lipid pairs versus the results under $F = G$ in chapter 4.

6.2. Parametric Bootstrap Null (PBN) Distribution Fitting under $\mu_F = \mu_G$

6.2.1. The Results from fad2 Dataset Using PBN

Following the bootstrap procedure described in section 6.1, the test statistic $tg*$, $SSD*$ and R_T^* from the bootstrap samples are fit to the four well-known distributions: Exponential, Gamma, Lognormal and Weibull. Table 6.2 shows the counts of K-S test statistics D from the 200 bootstrap samples. By the minimum D criterion from chapter 4, the exponential distribution is chosen to be the distribution class for tg, and Weibull distribution is chosen to be the distribution class for SSD and R_T statistics. Figure 6.2 uses the box plots to show the distribution of the K-S D statistic for the test statistics tg^*, SSD^* and R_T^*. As you can see that the chosen distribution of K-S D statistics in Figure 6.2 are consistent with the results from Table 6.2.

Table 6.2: The counts of the minimum K-S test statistic D_{Min} for the 4 candidate distributions in fad2 dataset.

	Distributions				
Statistics	**Exponential**	**Gamma**	**Lognormal**	**Weibull**	**Total**
tg^*	183	1	11	5	200
SSD^*	2	29	4	165	200
R_T^*	0	69	1	130	200

In the last row for R_T^*, among 200 minimum Ds in $D_{Min}^* = (D_{Min}^{*1}, D_{Min}^{*2}, ..., D_{Min}^{*b}, ..., D_{Min}^{*200})$, zero of them are from the Exponential distribution, 69 from the Gamma distribution, 1 from the Lognormal distribution, and 130 from the Weibull distribution. The Weibull distribution

outperforms all other three distributions for R_T^*. Hence, the Weibull distribution is chosen to be the null distribution for R_T^*.

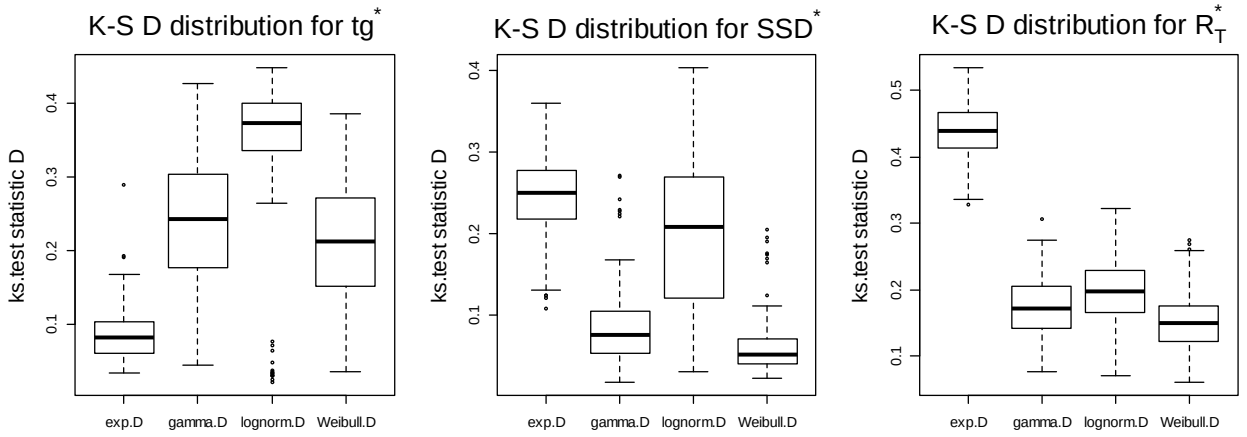

Figure 6.2: The distribution of K-S test statistic D for all three statistics tg*, SSD* and R_T^* in fad2

The box plots show the values of the Kolmogorov-Smirnov test (K-S test) D statistics from the 200 bootstrap samples for each of the 4 well-known distributions.

Table 6.3 shows the parametric null distributions for the three statistics from the chosen parametric distribution class. Three distributions are shown for each statistics. The parameters of the 5th percentile, mean and 95th percentile distributions are summarized from the 200 bootstrap fittings from the 200 MLEs. To be consistent with chapter 4, the 95th percentile bounding distribution would be used as the null distribution to get the list of findings. In Figure 6.3, the 95th percentile parametric null distribution $R_T \sim Weibull(7.88, 4.44)$ is overlaid with the data R_T in fad2. The p values for each test are computed from the following formula:

$$P(X > x_{observed}) = \int_x^{+\infty} f_{H_0}(x)dx,$$

where $x_{observed}$ = R_T is the observed statistics in the data and $f_{H_0}(x) = Weibull(7.88, 4.44)$. The local *fdr* multiple adjusting procedure is applied to control for multiple testing in the family of 4623 lipid pairs in fad2. There are 2159 lipid pairs in the final list of significant findings using PBN under the hypothesis $\mu_F = \mu_G$ in fad2.

Table 6.3: The final null distributions of the three statistics from the chosen parametric distribution with the parameter of estimates

The null distribution is shown in three forms, i.e., 5th percentile, mean and 95th percentile parametric distributions for each statistics. The final null distribution for the three statistics tg^*, SSD^* and R_T^* are the chosen distributions from the minimum K-S D criterion.

Statistics	5th percentile distribution	Mean distribution	95th percentile distribution
tg^*	*Exp* (0.50)	*Exp* (0.74)	*Exp* (1.1)
SSD^*	*Weibull*(1.36,0.57)	*Weibull*(2.11,1.00)	*Weibull* (2.89,1.43)
R_T^*	*Weibull*(2.83, 3.34)	*Weibull*(4.95, 3.84)	*Weibull* (7.88, 4.44)

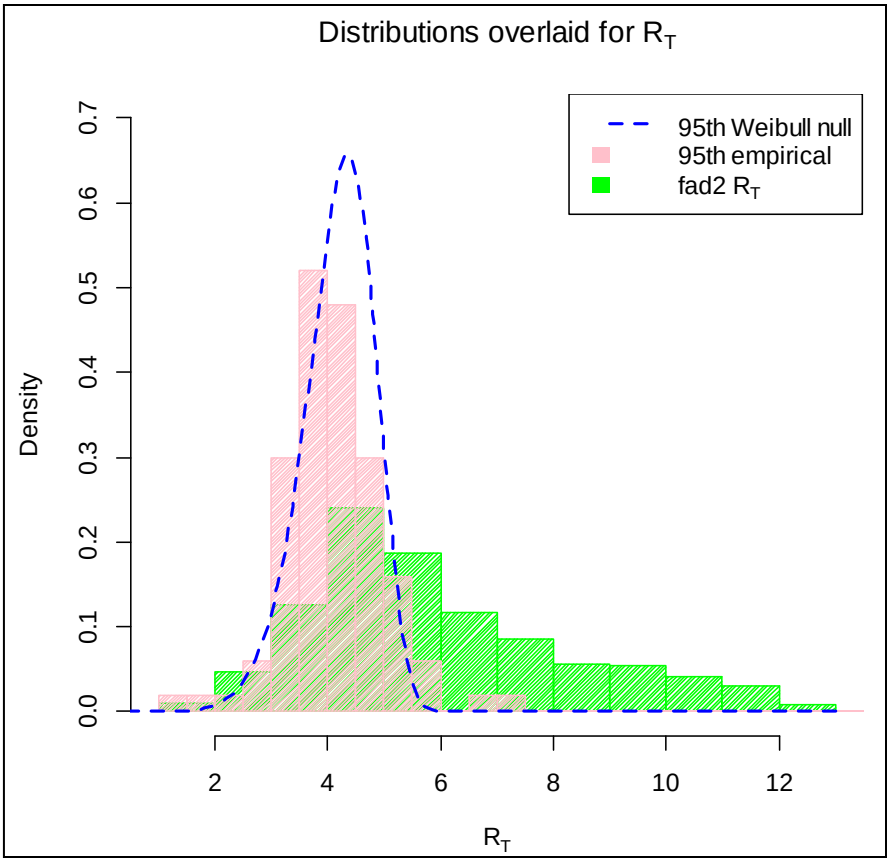

Figure 6.3: The 95th percentile empirical bootstrap distribution overlaid with the 95th percentile Weibull distribution and the real data from fad2

The pink histogram is the 95th empirical bootstrap null distribution. The blue dashed curve is the 95th percentile bounding distribution from the Weibull chosen distribution class. The green histogram is the distribution of R_T from fad2.

6.2.2. The Results from fad4 Dataset Using PBN

In a similar way as in section 6.2.1, the bootstrap samples under the null hypothesis $H_0 : \mu_F = \mu_G$ are too fitted to the four well-known distributions to find the distribution class. Table 6.4 lists the counts of the minimum K-S test statistics Ds. In the last row for R_T^*, among 200 minimum Ds in $D_{Min}^* = (D_{Min}^{*1}, D_{Min}^{*2}, ..., D_{Min}^{*b}, ..., D_{Min}^{*200})$, zero of them are from the Exponential distribution, 70 from the Gamma distribution, 2 from the Lognormal distribution, and 128 from the Weibull distribution. The Weibull distribution outperforms all other three distributions for R_T^*. Hence, the Weibull distribution is chosen to be the null distribution for R_T^*.

Figure 6.4 shows the distribution of the K-S test statistics D. The null distribution classes are exponential for tg, and Weibull for both SSD and R_T. The null distribution classes in fad4 are the same as those in fad2. This is an indication that the characteristics of the bootstrap samples are well captured by the PBN method regardless of what datasets we are exploring.

Table 6.5 gives the null distributions for 5th percentile, mean, and 95th percentile distributions for the three test statistics tg^*, SSD^* and R_T^* from the chosen distribution classes. To get the final results, tg^* and SSD^* distributions are ignored and the final results are derived from the 95th percentile distribution from R_T^* since R_T^* combines the advantage of the two statistics tg^* and SSD^* as specified in chapter 3. Figure 6.5 shows that the 95th percentile $R_T^* \sim Weibull(7.17, 4.49)$ can be used as the parametric null distribution to fit the 95th percentile bootstrap (pink) histogram to get a list of significant results in fad4 data.

Table 6.4: The counts of the minimum K-S test statistic D_{Min} for the 4 candidate distributions in fad4 under the assumption of $\mu_F = \mu_G$

	4 well-known distributions				
Statistics	**Exponential**	**Gamma**	**Lognormal**	**Weibull**	**Total**
tg^*	158	3	31	8	200
SSD^*	0	40	5	155	200
R_T^*	0	70	2	128	200

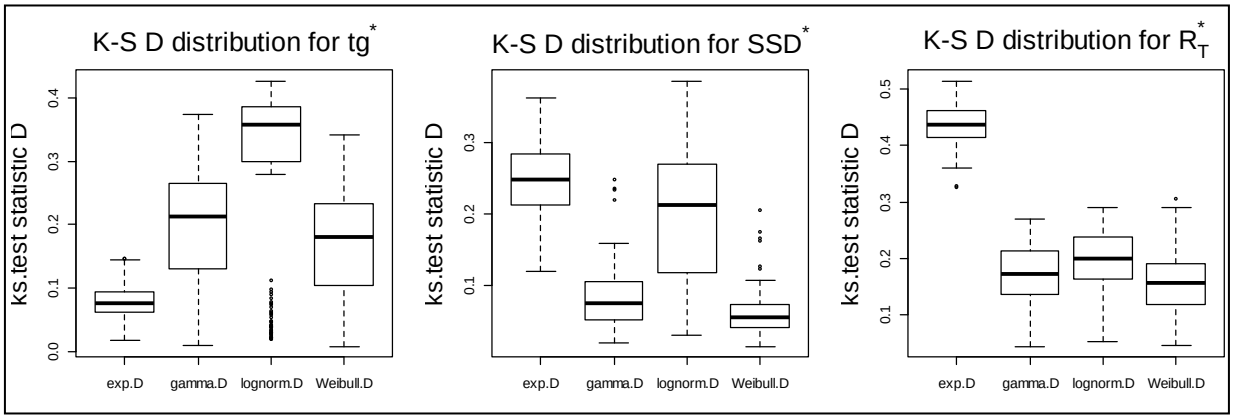

Figure 6.4: The distribution of K-S test statistic D for all three statistics tg*, SSD* and R_T^* in fad4

The box plots are from the real values of the Kolmogorov-Smirnov test (K-S test) D statistics from the 200 bootstrap samples for each of the 4 well-known distributions.

Table 6.5: The final null distributions of the three statistics from the chosen parametric distribution with the parameter of estimates in fad4

The null distribution is shown in three forms, i.e., 5th percentile, mean and 95th percentiles parametric distributions for each statistics. The final null distribution for the three statistics tg*, SSD* and R_T^* are the chosen distributions from the minimum K-S D criterion.

Statistics	5th percentile distribution	Mean distribution	95th percentile distribution
tg^*	*Exp* (0.54)	*Exp* (0.74)	*Exp* (1.10)
SSD^*	*Weibull*(1.41, 0.56)	*Weibull*(2.11, 1)	*Weibull* (3.09, 1.46)
R_T^*	*Weibull*(2.92, 3.38)	*Weibull*(4.79, 3.87)	*Weibull* (7.17, 4.49)

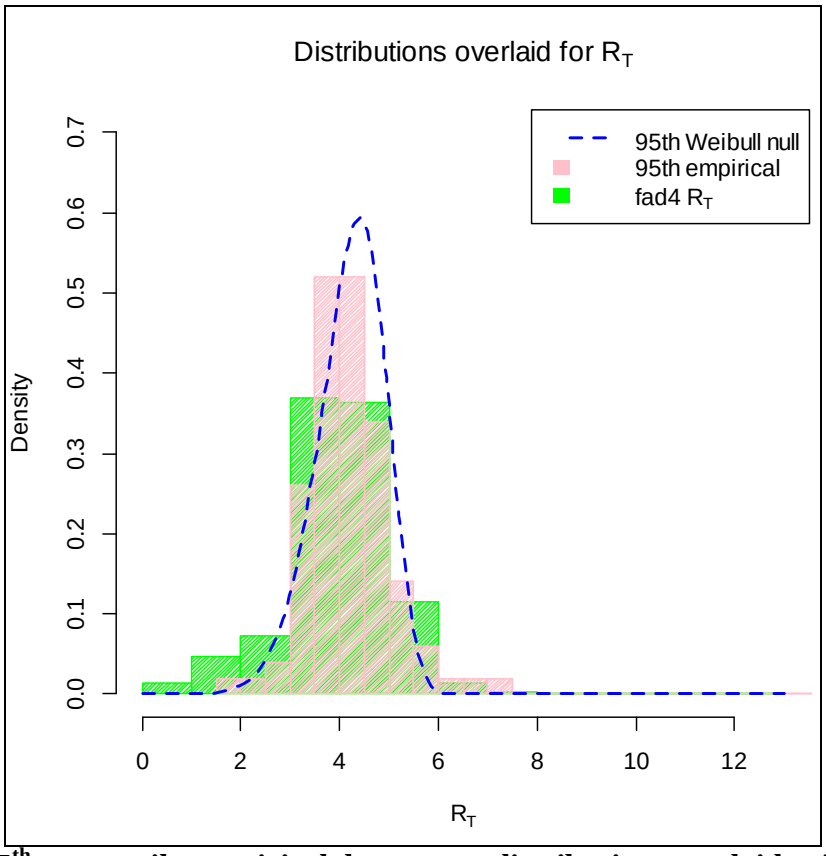

Figure 6.5: The 95th percentile empirical bootstrap distribution overlaid with the 95th percentile Weibull distribution and the real data from fad4

The pink histogram is the 95th empirical bootstrap null distribution. The blue dashed curve is the 95th percentile bounding distribution from the Weibull chosen distribution class. The green histogram is the distribution of R_T from fad4.

6.3. Mixture Normal Bootstrap Null (MNBN) distribution Fitting Under the Equal Mean Hypothesis

6.3.1. The Results from fad2 Dataset Using MNBN

The test statistics R_T^* from the bootstrap samples under the assumption $\mu_F = \mu_G$ are overlaid with the R_T distribution in fad2 in Figure 6.6. From Figure 6.6(c) we can see that there is still a large portion of the data to the right side of the null empirical distribution. Since the fad2 dataset showed strong mutation effect as shown in chapter 5 by the randomization test, a long list of findings are expected by using the MNBN method. It will be interesting to compare the results from the PBN and MNBN method in the strong signal dataset of fad2.

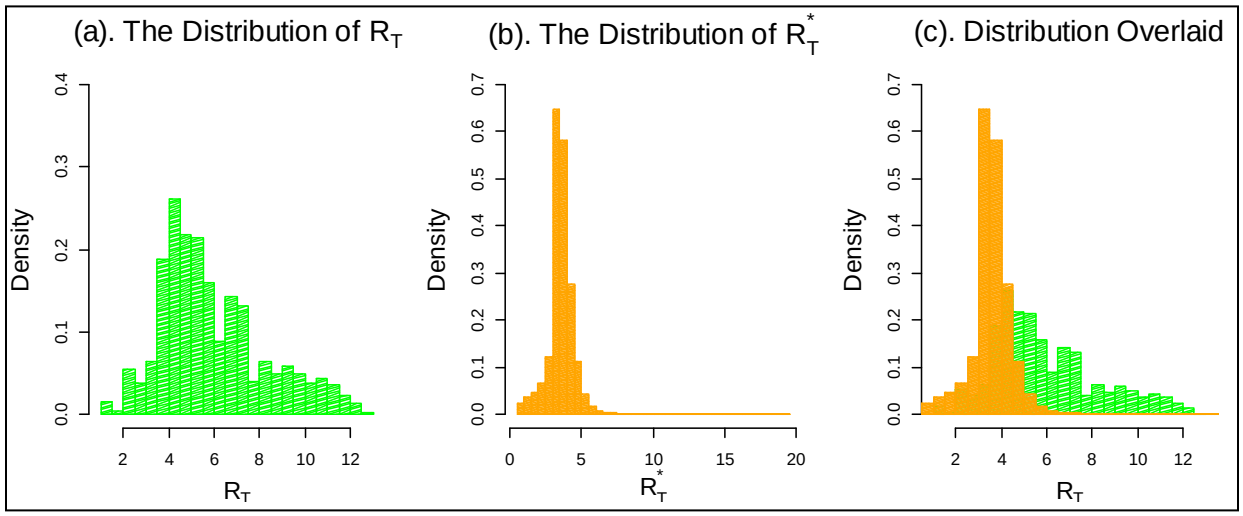

Figure 6.6: The R_T^* bootstrap empirical distribution overlaid with the R_T distribution for fad2 under the hypothesis of $\mu_F = \mu_G$

(a). the distribution of the R_T in fad2; (b). the empirical bootstrap distribution of R_T^*. (c). The bootstrap empirical distribution overlaid with the real data.

Two-component, three-component and four-component mixture normal distributions are fit to the bootstrap empirical distribution. To be consistent with the results from the previous chapters, a two-component mixture normal distribution is used to find the final list of findings. The MLEs and the 95% confidence intervals for the parameters are given in Table 6.6. Figure 6.7 shows the two-component mixture normal model to the empirical bootstrap distribution.

Table 6.6: The MLEs and the log-likelihood value for mixture normal null distribution in fad2 with two components and five parameters: π_1, μ_1, σ_1, μ_2, and σ_2.

Parameter π_1 is the proportion for the first component. μ_1 and μ_2 are the means for the two components. σ_1 and σ_2 are the standard deviations for both components.

Parameters	Estimates	Standard Error	95% confidence interval	Log-likelihood
π_1	0.70	0.000957	(0.703, 0.707)	-899496.6
μ_1	3.58	0.000654	(3.58, 3.582)	
μ_2	3.54	0.003752	(3.53, 3.545)	
σ_1	0.43	0.000752	(0.431, 0.434)	
σ_2	1.63	0.003074	(1.626, 1.638)	

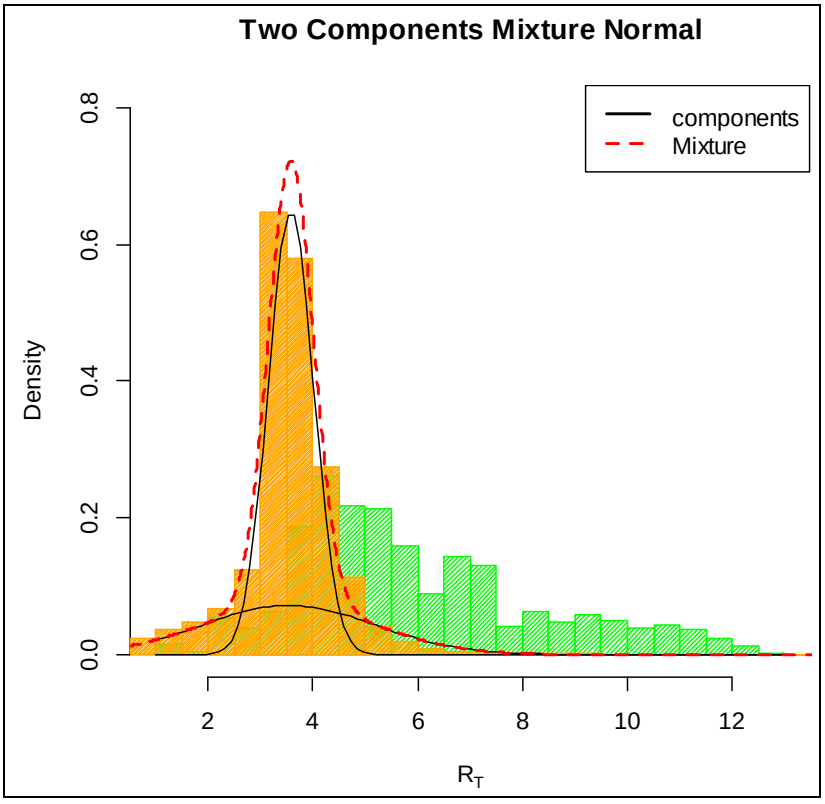

Figure 6.7: Two-component MNBN distribution of R_T^* in fad2
The two black solid curves are the two normal components. The red dashed curve is the mixture model from the two normal components.

The two-component $f_{H_0}(x) = 0.70 \cdot N(3.58, 0.43) + 0.30 \cdot N(3.54, 1.63)$ MNBN distribution in fad2 is utilized to find the p values for the tests. Since bigger R_T statistics show results, an upper-tail test is use to get a list of significant findings. The p values are calculated as

$$P(X > x_{observed}) = \int_x^{+\infty} f_{H_0}(x)dx = \int_x^{+\infty} \left[0.70 \cdot N(3.58, 0.43) + 0.30 \cdot N(3.54, 1.63)\right]dx,$$

where $x_{observed}$ = R_T is the observed statistics in the data. The local *fdr* multiple adjusting procedure is applied to control the type I error rate in the family of 4623 lipid pairs in fad2. 2643 lipid pairs are found to be significant by using the MNBN method under the null hypothesis H_0: $\mu_F = \mu_G$. In fad2, the final list is not shorter in using the MNBN method than the PBN method. Which method works better in the weak signal dataset in fad4? This will be investigated in the next section.

6.3.2. The Results from fad4 Dataset Using MNBN

The test statistics R_T^* from the bootstrap samples under the assumption $\mu_F = \mu_G$ are overlaid with the R_T distribution in fad4 in Figure 6.8. From Figure 6.8(c) we can see that there is a small portion of the data on the right side of the null empirical distribution.

To be consistent with the results from the previous chapters, two-component mixture normal distribution is used to find the final list of findings. The MLEs and the 95% confidence intervals for the parameters are shown in Table 6.7. Figure 6.9 shows the two-component mixture normal model fit to the empirical bootstrap distribution in fad4.

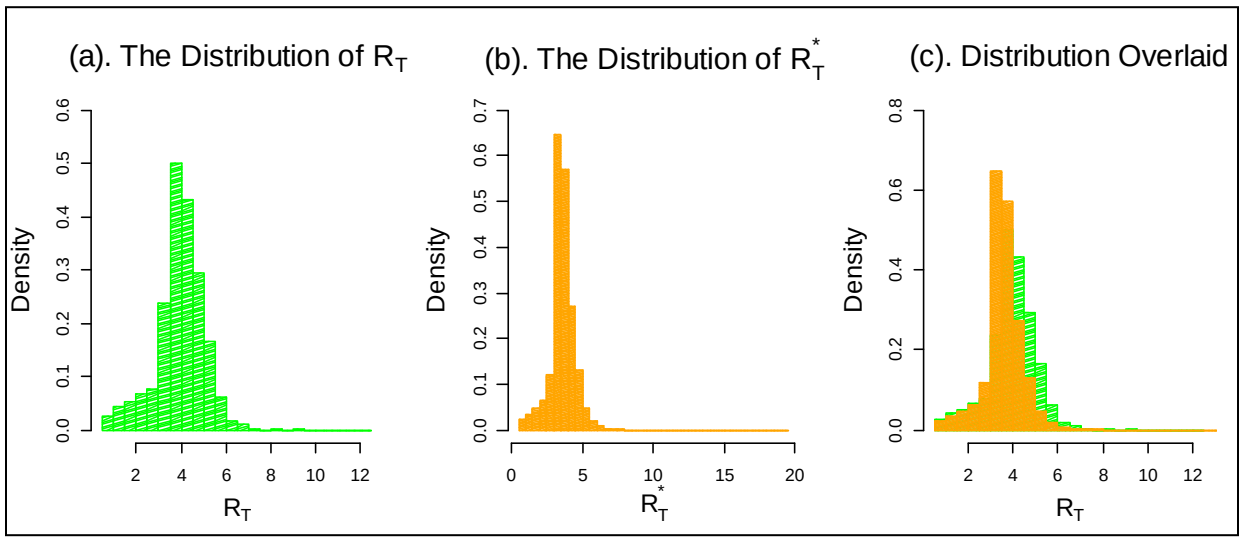

Figure 6.8: The R_T^* bootstrap empirical distribution overlaid with the R_T distribution for fad4 under the assumption of $\mu_F = \mu_G$

(a). the distribution of the R_T in fad4. (b). the empirical bootstrap distribution of R_T^*. (c). The bootstrap empirical distribution overlaid with the real data.

Table 6.7:The MLEs and the log-likelihood value for mixture normal null distribution in fad4 with two components and five parameters $\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2$.

Parameter π_1 is the proportion for the first component. μ_1 and μ_2 are the means for the two components. σ_1 and σ_2 are the standard deviations for both components.

Parameters	Estimates	Standard Error	95% confidence interval	Log-likelihood
π_1	0.64	0.001296	(0.359, 0.364)	
μ_1	3.57	0.002962	(3.58, 3.59)	
μ_2	3.58	0.00075	(3.571, 3.573)	-795694.8
σ_1	0.40	0.002409	(1.433, 1.442)	
σ_2	1.44	0.000767	(0.402, 0.405)	

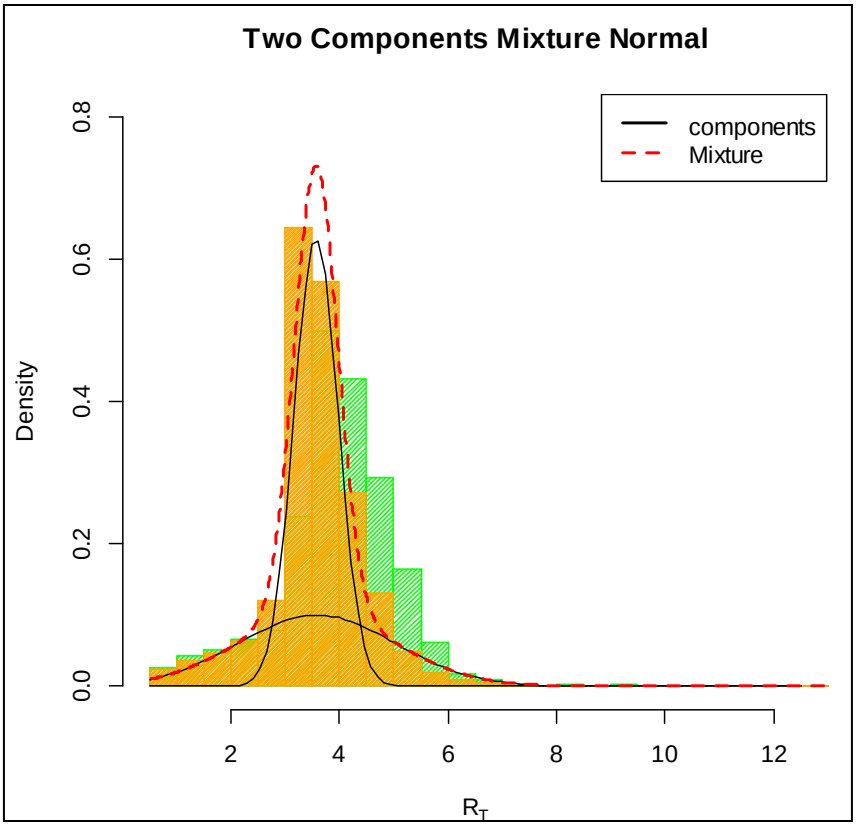

Figure 6.9: Two-component MNBN distribution of R_T^* in fad4 data

The two black solid curves are the two normal components. The red dashed curve is the mixture model from the two normal components.

The two-component mixture $f_{H_0}(x) = 0.64 \cdot N(3.57, 0.40) + 0.36 \cdot N(3.58, 1.44)$ MNBN distribution in fad4 is utilized to find the p values for the tests. The p values are calculated as

$$P(X > x_{observed}) = \int_x^{+\infty} f_{H_0}(x)dx = \int_x^{+\infty} [0.64 \cdot N(3.57, 0.40) + 0.36 \cdot N(3.58, 1.44)]dx .$$

where $x_{observed}$ = R_T are the observed statistics in the data. The local *fdr* multiple adjusting procedure is applied to control the type I error rate in the family of 2964 lipid pairs in fad4. Only 6 lipid pairs are found to be significant by using the MNBN method under the null hypothesis H_0: $\mu_F = \mu_G$.

6.4. Discussion

After the exploration of the methods using PBN and MNBN under two null hypotheses $H_0 : F = G$ in chapter 4 and chapter 5, respectively, and $H_0 : \mu_F = \mu_G$ in chapter 6, a summary of methods is shown in Table 6.8. The PBN results were produced under the 50th percentile parametric Weibull null distribution for R_T statistics in each dataset. The MNBN results were generated by using the two-component mixture normal distributions from a large vector of R_T statistics.

Table 6.8: Comparison of the results using PBN and MNBN methods under the null hypothesis $\mu_F = \mu_G$ and $F = G$

The results from PBN method are produced using the 50th percentile Weibull null distributions for R_T statistics. "Prop. of findings" represents the number of the significant findings in the total lipid pairs satisfying the y = 2 screening criteria.

Null hypotheses	Methods	PBN		MNBN	
	Datasets	**fad2**	**fad4**	**fad2**	**fad4**
	Total number of lipid pair	4623	2964	4623	2964
$\mu_F = \mu_G$	# of Significant findings	2623	92	2643	6
	Prop. of the findings	0.57	0.031	0.57	0.002
$F = G$	# of Significant findings	3207	551	2626	10
	Prop. of the findings	0.69	0.19	0.57	0.0034

In Table 6.8, the proportion of the findings indicate the proportion of significant pairs in the total number of lipid pair in the datasets fad2 and fad4 when the screening criteria y = 2 is satisfied. We can see that the proportions of the significant findings in the two methods under the two different null hypotheses are both very close to the mutation effects shown in Table 5.1. At this point, it is hard to conclude that one method is better than the other. Also, the two

different null hypotheses do not show a substantial difference in the performance of the methods. To get a conclusion on the performance of the methods, a further extensive simulation study should be considered. However, the MNBN distribution with two components has five parameters and the Weibull PBN distribution has two parameters. The MNBN distribution may have more flexibility to capture the shape of the bootstrap samples than the Weibull PBN distribution. Hence, the MNBN method will be used in chapter 9 when proposing a method to simulate realistic data and evaluate the properties of the simulated data.

Summary remarks

In this chapter, we have used the Mixture Normal Bootstrap Null (MNBN) distribution and the Parametric Bootstrap Null (PBN) distribution to find a list of significant lipid pairs. Both MNBN and PBN are used to simulate the null distribution that is then fitted to the empirical bootstrap distribution . We suggest that the MNBN distribution may have more flexibility to capture the shape of the empirical bootstrap distribution than PBN. But in general the methods perform similarly under the two different null hypotheses $\mu_F = \mu_G$ and F = G.

Chapter 7 - A Mixture Model to Fit the R_T Distribution in the Data

7.1. Normal Mixture Distributions for R_T

As shown in chapter 4, the bootstrap could be useful for finding a null distribution for a statistic whose distribution cannot be not readily obtained. However, it should be noted that the bootstrap technique requires resampling under a specified null hypothesis and that the specified null hypothesis might be too restrictive for the current application of targeted lipid analysis in a WT-MT experiment. Efron (2004) proposed a method for finding an empirical null distribution for a standard normal test statistic (here referred to as a z-score). A distribution of z-scores was obtained from tests of multiple hypotheses (if a test other than a normal distribution based test was done, the test statistic or p-value was transformed to a corresponding normal z-score). Using a combination of theory and some heuristics, Efron (2004) derived an empirical null distribution, one that was not necessarily the standard normal. Then, z-scores in the tails of this distribution that were deemed atypical were seen as evidence of true discoveries. An approach analogous to Efron's (but for now less rigorously developed) is considered in this chapter, yielding some interesting results.

Looking back at the distribution of the R_T statistics from the fad2 dataset, one notices a pattern of an apparent bimodal distribution. Since large values of R_T are considered most interesting, intuition suggests that there may be a distribution of lower values of R_T that represent the empirical null cases (i.e., a pair is not affected by the mutation). A component distribution of larger values of R_T may then be considered the "interesting cases". In fact, the bimodal shape suggests that a mixture of normal distributions may be sufficiently flexible for modeling the entire distribution of R_T statistics. Rather than a bootstrap null distribution as was done in earlier chapters, in this chapter, the following questions will be addressed: (1). How many normal components are needed in a normal mixture model? (2). What are the parameter estimates and the confidence intervals for the parameters in the mixture model, and how can they be interpreted? (3). Using the normal mixture model, what are the posterior probabilities that, given the value of R_T for each lipid pair, it is in fact a pair that is significantly affected by the mutation? (4). What is the proportion of significant pairs?

The focus is on R_T because this statistic includes information on both metrics, tg and SSD. Let $x = R_T$. The probability density function for a normal distribution is expressed by

equation (5.1). The probability density function of the mixture normal distribution has similar form in (2.5) with k normal distributions as its components. Note that the mixture normal model approach in this chapter is similar with the Mixture Normal Bootstrap Null (MNBN) distribution which was introduced in chapter 5. The main focus of the MNBN is to find a null distribution using the mixture normal distribution. In this chapter, the mixture normal distribution would be used to fit the real data and use one of the components from the mixture to model the data under the alternative hypothesis (i.e. a pair is affected by the mutation). The mixture model will be used to fit the data using maximum likelihood estimation (MLE). Again, we are assuming that the roughly 4,600 lipid pairs satisfying the screening for $y = 2$ are mutually independent. This assumption is likely not met in a lipidomics experiment. Nevertheless, it has been made in many high-dimensional "omics" studies for the purpose of fitting a mixture distribution (some of these cited in Gadbury at al., 2008) as discussed in chapter 5. The fitted mixture distribution is still a representation of "relative model fit" in the sense of an expected fit over many realizations from the experiment. Correlation is likely to make results more variable from study to study. The independence assumption will be discussed further in chapter 10 when future directions are discussed.

Let i index lipid pairs in the fad2 dataset, where $i = 1, 2, \ldots, M$, where M is around 4600, depending on the dataset. The research interest is H_{0i} versus H_{ai}. In the H_{0i}, the i^{th} reactant and product pair is not affected by the mutation, and in the H_{ai}, the i^{th} reactant and product pair is affected by the mutation.

The likelihood function for the R_T statistic from a normal mixture model with $k = \nu + 1$ components can be expressed as in equation (5.2), where x_i is the R_T statistic for the i^{th} test. The probability density function of the normal mixture model can be expressed as

$$f(x_i / \mu_j, \sigma_j) = \sum_{j=1}^{k} \pi_j \cdot f_j(x_i / \mu_j, \sigma_j), \tag{7.1}$$

Where the subscript j stands for the j^{th} component in the mixture model. π_j s are the mixing proportions on each of the component densities, satisfying the constraints $\sum_{j=1}^{k} \pi_j = 1$, and $0 \le \pi_j \le 1$.

The most common mixture model in high-dimensional analysis tends to be a two component mixture model. One component is intended to model test statistics (or p-values) from tests for which the null hypothesis is true, and the other for tests for which it is false. Suppose that the mixture model with two normal components is used. The probability density function for a normal mixture model with two components is

$$f(x_i / \pi_0, \mu_0, \sigma_0, \mu_1, \sigma_1) = \pi_0 \cdot f_0(x_i / \mu_0, \sigma_0) + \pi_1 \cdot f_1(x_i / \mu_1, \sigma_1),\ i = 1,\mathrm{K}\ , M\ , \qquad (7.2)$$

where π_0 = Pr (the lipid pair is not affected)

π_1 =1- π_0 = Pr (the lipid pair is affected)

$f_0(x_i / \mu_0, \sigma_0)$ = the density of x_i (i.e. R_{Ti}) under the null hypothesis if the lipid pair is not affected by the mutation. It has the form of (5.1) with the mean μ_0 and standard deviation σ_0.

$f_1(x_i / \mu_1, \sigma_1)$ = the density of x_i under the alternative hypothesis if the lipid pair is affected by the mutation. It has the form of (5.1) with the mean μ_1 and standard deviation σ_1.

By Bayes' rule, the posterior probability of a lipid pair at a given x_i = R_{Ti} that is affected by the mutation is defined as

$$P(affected / x_i) = 1 - \frac{\pi_0 \cdot P(x_i / unaffected)}{P(x_i)} = 1 - \frac{\pi_0 \cdot f_0(x_i / \mu_0, \sigma_0)}{f(x_i / \pi_0, \mu_0, \sigma_0, \mu_1, \sigma_1)} \qquad (7.3)$$

so that one minus the quantity in (7.3) is the posterior probability of a lipid pair that is unaffected by the mutation, which is the local false discovery rate (*lfdr*) defined by Efron (2004).

The maximum likelihood estimates, MLEs, of the parameters π_j, μ_j, and σ_j can be obtained by using the EM algorithm for a normal mixture model that maximizes the conditional expected complete-data log-likelihood at each M-step of the algorithm (McLachlan and Peel 2000; Meng and Rubin 1993). The standard error can be obtained from the bootstrap samples in order to compute large sample confidence intervals for the population parameters in the model. The biologically functional lipid pairs which were introduced in chapter 3 will be used as an indication of some thresholds for R_T for a specific lipid pair to evaluate whether it is significantly affected by the mutated genes.

7.2. Test the Number of Components in the Normal Mixture Models

Figure 7.1 shows the two-component mixture normal distributions fitting to the R_T distribution in all fad datasets. Figure 7.2 shows the three-component mixture normal models in each dataset.

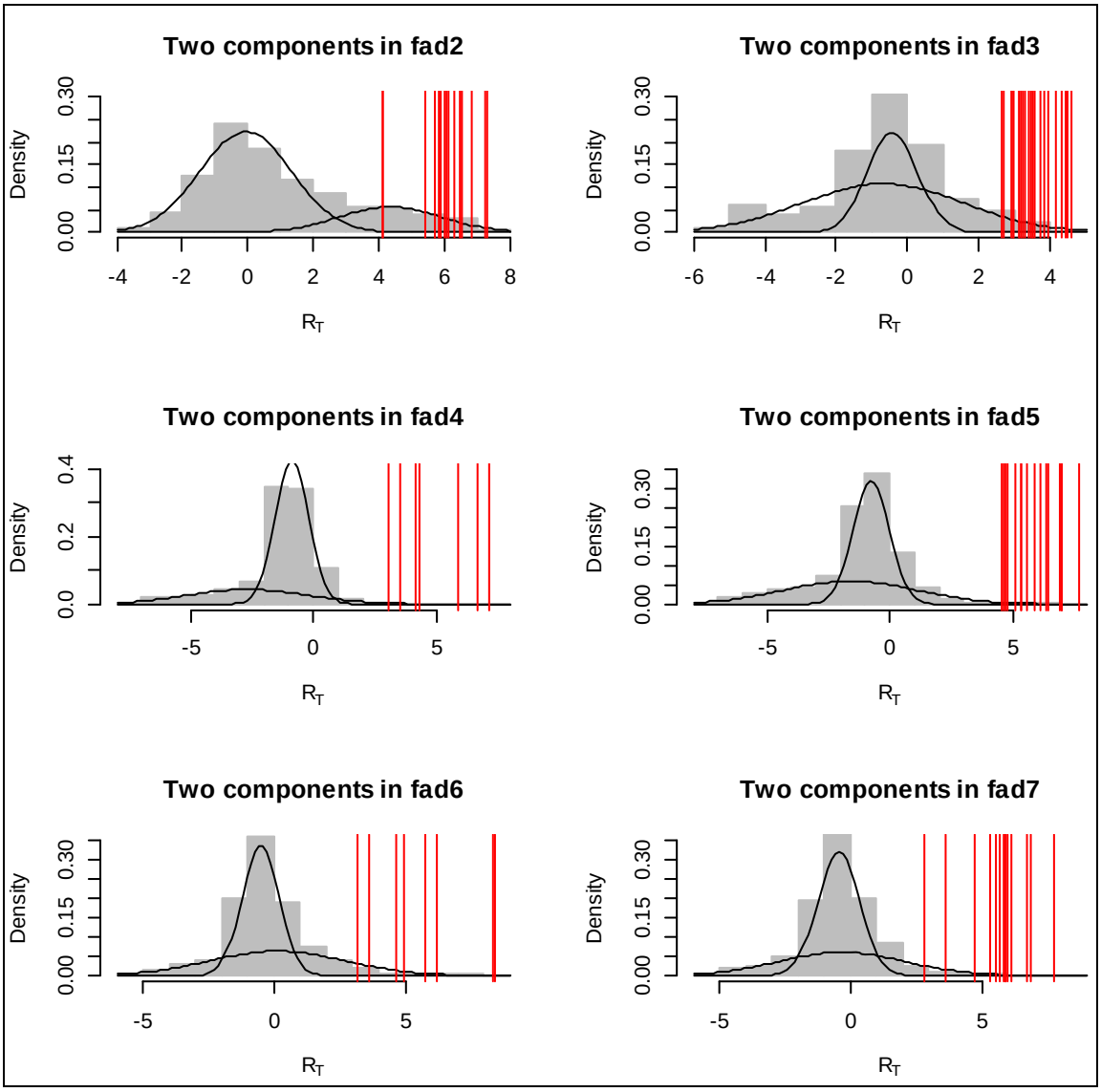

Figure 7.1: The two-component mixture normal distributions fitting to all the fads datasets The gray histograms are the R_T distribution in each dataset. The two black curves are the two normal components. The red vertical lines are the real R_T statistics for the biologically functional lipid pairs in each dataset.

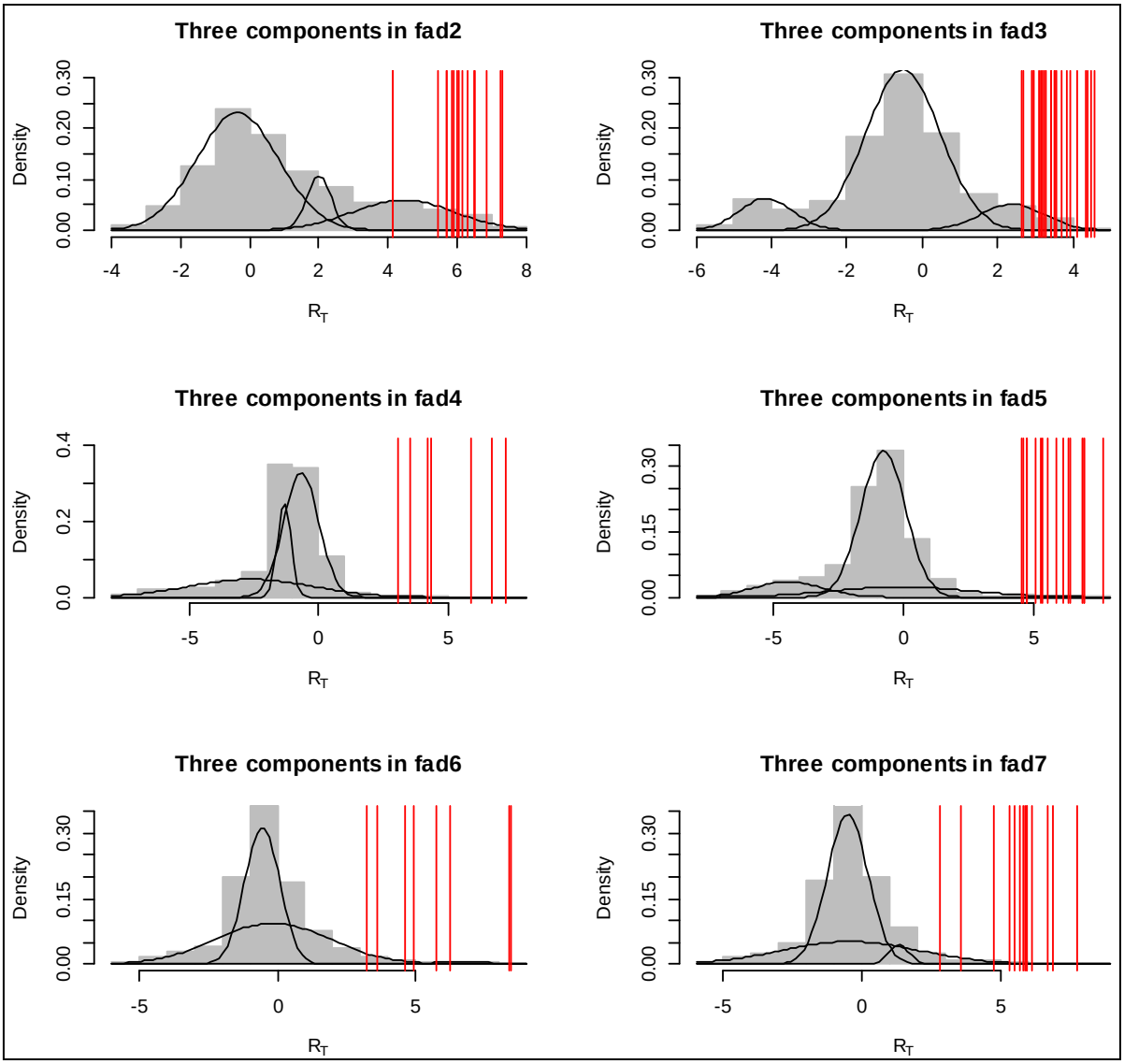

Figure 7.2: The three-component mixture normal distributions fitting to all the fads datasets

The gray histograms show the R_T distribution in each dataset. The three black curves are the three normal components. The red vertical lines stand for the statistics R_T from the biologically functional lipid pairs.

Observing the two-component mixture model for fad2 in Figure 7.1, the two modes can be modeled by two well-separated normal components. It will be assumed that the normal component on the right-hand side models data for biologically functional lipid pairs (red vertical lines) and it will be used as the alternative hypothesis to get the final findings. The component on the left-hand side will be used as the null hypothesis model for the lipid pairs. In fad2 in Figure 7.2, a three-component normal mixture model is used to fit the distribution of the statistics R_T.

The component on the most right-hand side will be assumed to be the alternative model. The other two components then represent the null model. From the two- and three-component mixture models in both Figures 7.1 and 7.2, it seems that fad2 can be modeled by both two- and three-component mixture normal distributions because there is a clear separation between the null and the alternative model in each figure. On the other hand, for fad3 the three-component mixture model seems to be better than the two-component model, because in the three-component mixture model the separation between the nulls and the alternative model is clearer. For the rest of the datasets, this method may not be applicable because neither two- nor three-component mixture model can give a clear separation between the null and the alternative since there is insufficient density in the right tail of the distribution to identify a distinct normal component distribution.

The fad2 dataset will be used as an example to illustrate the mixture model method in this chapter. A initial question regards the number of components that should be used in the normal mixture model.

Bootstrap algorithm for testing the number of components

1. Obtain the Maximum likelihood Estimates and the log-likelihood function from equation (5.2) for K components and $K-1$ components mixture normal models, where K = 1, 2, … k.
2. Let loglik(k) and loglik(k-1) be the log-likelihood values for the mixture normal distribution with K components and with $K-1$ components, respectively. Define the test statistics T_{obs} = loglik(k) – loglik(k-1), where T_{obs} is the observed test statistic.
3. Simulate data from a $K-1$ component mixture model using the MLEs from step 1.
4. Fit those data using both $K-1$ and K-component normal mixture models. Extract the log-likelihood values loglik(k*-1) and loglik(k*).
5. Calculate the test statistic T* = loglik(k*) − loglik(k*-1) from models fit to the simulated data.
6. Let B = 1000. Repeat steps 3 to 5 1000 times. Record the T*s from each loop to get a vector $T^* = (T_1^*, T_2^*, ..., T_{1000}^*)$.
7. Test the hypotheses $H_0 : K-1$ components versus $H_A : K$ components by using the p value which is calculated with $\frac{\#(T^* > T_{obs})}{1000}$.

Note that bootstrap samples were generated under the null hypothesis with $K-1$ components. In the fad2 dataset, the hypotheses H_0: Two components versus H_A: Three components was tested with a p value equal to 0.038. Thus, the three-component mixture normal is selected. Then, H_0: Three components versus H_A: four components was tested with a p value equal to 0.183. We conclude that the three-component mixture normal is an adequate fit versus the four component normal mixture. In the three-component mixture model, the component on the right-hand side is modeling the larger values of R_T and will be used as a component distribution to represent the significant findings.

7.3. Estimation and Confidence Intervals for the Parameters in Normal Mixture Models

Three-component normal mixture models in fad2 is used here to illustrate the method. In the log-likelihood function in (5.2), a normal mixture model for the statistics R_T with three components has eight parameters, π_1, π_2, μ_1, μ_2, μ_3, σ_1, σ_2 and σ_3, as defined in (7.1). The statistical software R package "mixtools" is used to complete the task of the EM algorithm. (http://cran.r-project.org/web/packages/mixtools/mixtools.pdf). The EM algorithm for normal mixture model maximizes the conditional expected log-likelihood at each M-step of the algorithm (McLachlan and Peel 2000). The MLEs are found after the convergence is declared at a maximized log-likelihood value. Meng and Rubin (1993) developed an ECM algorithm to add one more extra E-step in between the E- and M-step to update the estimates. In this extra conditional maximization E- step, the iteration can update the means conditional on variances or update the variances conditional on the means. Table 7.1 shows the MLEs for three-component normal parameters and the 95% confidence intervals. The log-likelihood values is -9940.318 from the three-component model. The standard errors of the estimates can be obtained in the "mixtools" R package which utilized a bootstrap procedure for the specified mixture model.

Table 7.1 shows that three confidence intervals for μ_1, μ_2, and μ_3 are (-0.421, -0.29), (1.947, 2.060) and (4.252, 4.641). Note that the means of the three components do not overlap, and their densities have separation. Also, the confidence intervals for σ_1, σ_2 and σ_3 are (1.139, 1.234), (0.289, 0.406) and (1.342, 1.585). The confidence intervals for the standard deviations did not overlap either. Figure 7.3 shows the three component normal density

curves (in black solid curves) and the three-component normal mixture curve (in red dashed line).

Table 7.1: The confidence intervals for the parameters in a normal mixture with three components including eight parameters π_1, π_2, μ_1, μ_2, μ_3, σ_1, σ_2 and σ_3.

Parameter π_j is the mixing proportion for the j^{th} component. μ_j and σ_j are the means for the j^{th} components. The log-likelihood value is produced by using the MLEs from this table.

Parameter	Estimates	Standard Error	95% CI	Log-likelihood
π_1	0.689	0.011	(0.666, 0.711)	-9940.318
π_2	0.093	0.010	(0.073, 0.113)	
μ_1	-0.356	0.033	(-0.421, -0.29)	
μ_2	2.003	0.029	(1.947, 2.060)	
μ_3	4.447	0.099	(4.252, 4.641)	
σ_1	1.191	0.027	(1.139, 1.234)	
σ_2	0.347	0.030	(0.289, 0.406)	
σ_3	1.463	0.062	(1.342, 1.585)	

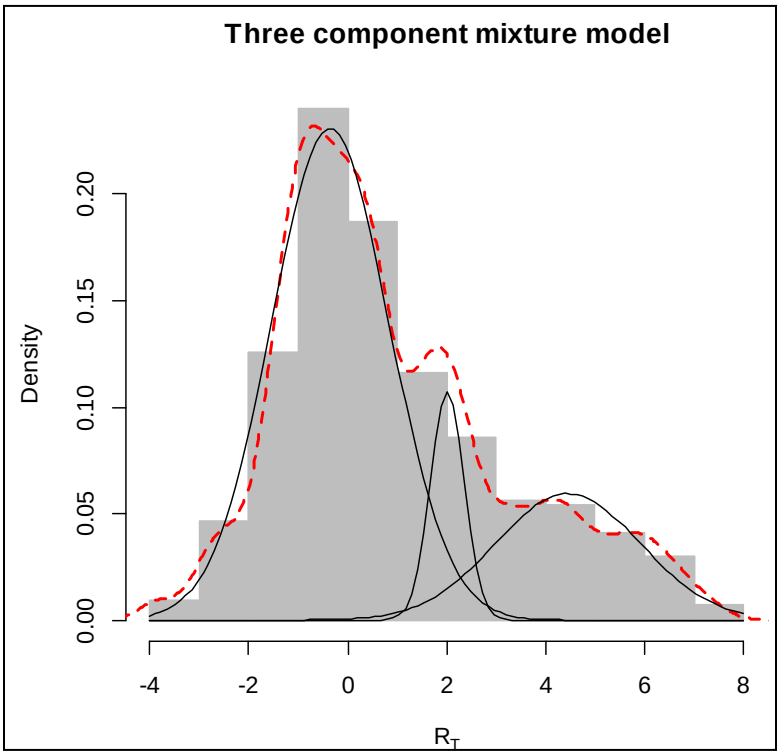

Figure 7.3: The mixture model fit of R_T in fad2 with three normal components

The red dashed curve is the mixture model distribution. The black solid curves are the three normal components with the estimated parameters shown in Table 7.1.

From Table 7.1 and Figure 7.3 we can see that the first normal component $N(-0.3, 1.191)$ on the left models most of the data with an estimated proportion of 0.689. The second component $N(2.003, 0.347)$ models the least amount of the data with an estimated proportion of 0.093. The third normal component $N(4.447, 1.463)$ models a moderate amount of the data with a proportion of 0.218. The mixture model with three components is:

$$f(x) = 0.689 \cdot N(-0.3, 1.191) + 0.093 \cdot N(2.003, 0.347) + 0.218 \cdot N(4.447, 1.463) \qquad (7.4)$$

The normal component $N(4.447, 1.463)$ will model the larger values of R_T and will be used as the alterative model to find the interesting lipid pairs. The other two normal components $N(-0.3, 1.191)$ and $N(2.003, 0.347)$ will be used as null components to model the lipid pairs with no mutation effects. The posterior probabilities for the interesting cases are then computed from (7.3) except that the null density comprises two components.

7.4 The Results for fad2 Dataset by Using Three-component Normal Mixture

The posterior probabilities are computed for each lipid pair in fad2 by using the mixture normal distribution with three components. There are 17 biologically functional lipid pairs in fad2. We want to know how many lipid pairs appear as significant results according to the posterior probabilities of each biologically functional pair.

Figure 7.4 shows the posterior probabilities for all 4623 lipid pairs that satisfy the screening scheme $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ in fad2. In Figure 7.4, the black solid curve is the posterior probabilities for each lipid pair at their R_T statistic from the three-component normal mixture model. The 17 biologically functional lipid pairs posterior probabilities are used as 17 different cutoff points above which "significant results" are determined.

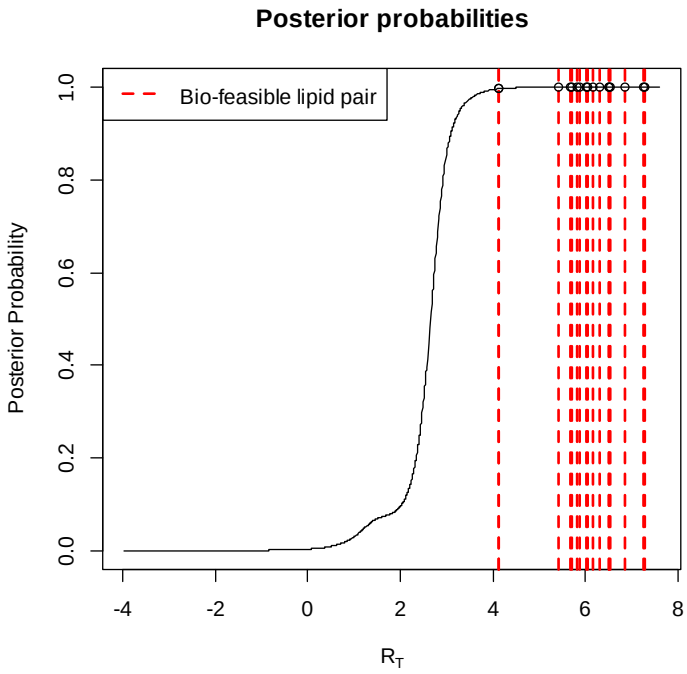

Figure 7.4: The posterior probabilities of the R_T for each lipid pair in fad2
The red dashed vertical lines are the biologically functional lipid pairs R_T statistics. The solid black curve is the posterior probabilities from the three-component mixture.

Table 7.2 shows the significant results by using 17 different cutoff points. The 17 different cutoff points are from the corresponding 17 biologically functional pair posterior probabilities. In table 7.2 the 17 cutoff points are listed in a descending order of their posterior probabilities. For example, the first pair, namely PC34_1_PC34_2, has the largest (biggest posterior probability) cutoff point of 0.999 with a R_T value of 7.28. Among all 4623 lipid pairs, 14 lipid pairs are significant with a proportion of 0.00303 of the total. Among the 14 significant lipid pairs, there are only 6 distinct reactants appearing in the results.

In the last row of Table 7.2, the biologically functional lipid pairs PC36_2_PC36_3 has a posterior probability of 0.94. This R_T value, 4.135, is the smallest among all 17 biologically functional R_T values. Its posterior probability 0.94 is also the smallest. If this posterior probability is used as the cutoff point, the proportion of significant pairs is 0.12 out of all 4623 lipid pairs in fad2. The number of significant lipid pairs is 566 with 25 distinct reactants appearing in the 566 significant results.

There are in total 67 possible reactants in the fad2 dataset that has a total of 4623 lipid pairs satisfying the screening criteria for $y = 2$. The 25 distinct reactants from the significant lipid pairs in the last row of Table 7.2 are listed in Table 7.3.

Figure 7.5 shows the proportions of the significant pairs out of a total 4623 ($y = 2$) lipid pairs and the proportion of significant distinct reactants. Those two plots are produced from the last two columns in Table 7.2. Comparing Figure 7.5 (a) and (b), one can see that the two proportions both decrease dramatically as R_T increases. This means in the significant results, the proportion of significant pairs can be decreased if we use a more restrictive cutoff points (posterior probability).

Table 7.2: The results for the significant lipid pairs using 17 biologically functional lipid pairs from three-component mixture model

Use *i* as the index for the 17 biologically functional lipid pairs. Column 1 is the number of the 17 biologically functional lipid pairs. Column 2 shows the names of those biologically functional lipid pairs. Column 3 stands for the i^{th} R_T values for the i^{th} biological pair. Column 4 shows their corresponding posterior probabilities. This whole table is sorted with the posterior probabilities in column 4 in a descending order. In column "Proportion", the i^{th} proportion stands for the proportion of significant pairs using the i^{th} posterior probability as the cutoff point. Column "Number sig.pairs" shows the number of significant pairs by using the i^{th} posterior probability as the cutoff point. Column "Number distinct.reactant" shows the number of distinct reactants.

Number	AB.name	R_T	Posterior Prob	Proportion	Number sig.pairs	Number distinct.reactant
1	**PC34_1_PC34_2**	7.280	0.99999997	0.003028	14	6
2	**PC36_2_PC34_2**	7.255	0.999999967	0.003245	15	6
3	**PC36_2_PC38_5**	6.858	0.999999837	0.009518	44	10
4	**PC34_1_PC38_3**	6.522	0.999999393	0.01817	84	14
5	**PC36_2_PC38_3**	6.507	0.999999355	0.018603	86	14
6	**PC36_2_LysoPC18_2**	6.503	0.999999347	0.018819	87	14
7	**PE36_2_LysoPE18_2**	6.310	0.999998623	0.025525	118	15
8	**PC38_2_PC38_3**	6.151	0.999997477	0.031149	144	16
9	**PE34_1_PE40_2**	6.062	0.999996464	0.035258	163	16
10	**PE36_2_PE40_2**	6.014	0.999995768	0.036989	171	16
11	**PE34_1_LysoPE16_0**	5.873	0.999992843	0.041964	194	17
12	**PE36_2_LysoPE16_0**	5.830	0.9999916	0.044992	208	17
13	**PC34_1_PC40_4**	5.698	0.999986278	0.051482	238	17
14	**PC36_2_PC40_4**	5.688	0.999985784	0.051914	240	17
15	**PE36_2_PE38_5**	5.426	0.999962949	0.061865	286	17
16	**PC34_1_PC36_3**	4.139	0.99680312	0.122215	565	25
17	**PC36_2_PC36_3**	4.135	0.996757692	0.122431	566	25

Table 7.3: The 25 distinct reactants from the significant lipid pairs using the lowest posterior probability for the biological pair PC36_2_PC36_3 from the last row of Table 7.4
In this table, the order of the significant pair is by columns.

PE34_1	**PS42_1**	**PC38_2**	**PG34_1**	**PS36_4**
PC34_1	**PS40_1**	**PI36_2**	**PE36_1**	**PE42_4**
PC36_2	**PS34_1**	**PE34_4**	**PS38_1**	**PS36_1**
PE36_2	**LysoPC18_1**	**PI36_1**	**PG32_1**	**PS36_2**
PI34_1	**LysoPE18_1**	**DGDG34_1**	**DGDG36_4**	**PA36_2**

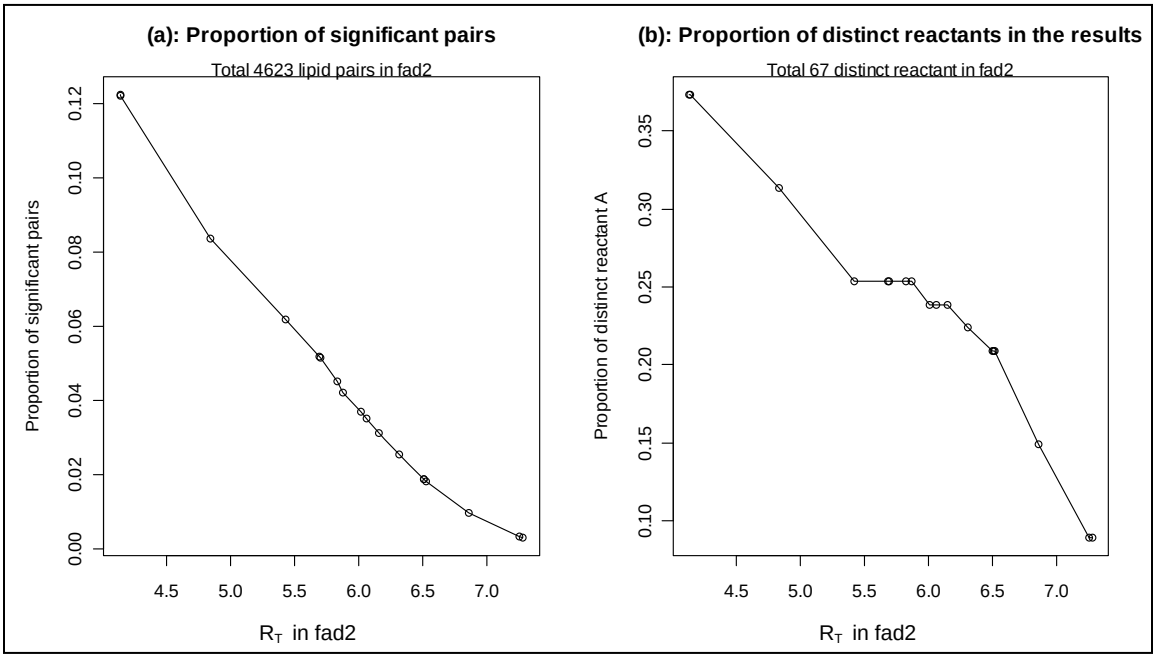

Figure 7.5: The proportion of significant pairs and the proportion of significant distinct reactants in fad2
(a) The proportion of significant pairs among 4623 pairs at 17 different cutoff points. The 17 cutoff points are the posterior probabilities of the biologically functional lipid pairs. (b) The proportions of number of significant distinct reactants from their corresponding significant results from panel (a). The total number of possible reactants is 67 in all 4623 pairs.

Some summary remarks

In conclusion, the normal mixture model method appears to be a useful approach for identifying reactant-product lipid pairs that are significantly affected by the mutated genes in some datasets. Previous key assumptions need more thorough consideration. The first is that the distribution of R_T includes component distributions for the null results and results that are "true findings". While this approach is consistent with other approaches commonly used for genetic expression data, some additional justification will be needed to characterize the "empirical null"

distribution for R_T and the degree to which a fitted normal mixture component is valid. This could involve considering the structure of the R_T statistic as a function of the sample means, and then determining the distributional characteristics of the statistic. This is somewhat similar to the work that Efron (2004, 2007) did for an empirical distribution of a test statistics under the null. Another issue is the likely strong dependence among pairs. The degree to which dependence may affect the results and whether there should be some adjustments for this need more consideration. This is similar to Efron's (2007) adjustment to estimated FDRs for dependence among gene expression data (across genes).

Chapter 8 - ANOVA Approach for the Pathway Analysis

8.1. Two-way ANOVA Model with an Interaction Term

In chapter 3 we explored the relations between a reactant A and a product B in the lipidomic data. The relations of the means in the lipid pair have to satisfy the screening scheme $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ in order for them to be a candidate reactant-product pair on the pathway. After consideration of the evidence in the data for pairs of lipids to be a candidate reactant-product pair affected by the mutation, it became clear that a simple ANOVA approach could be useful as well. This is explored here. Let z_{ijk} be the concentration of a lipid. Index i = 1, 2, denotes the lipid in a possible reactant-product pair. Index j = 1, 2, denotes the treatment groups WT and MT. Index k = 1, 2, …, n denotes the samples in the j^{th} treatment within the i^{th} role of lipid, and n denotes the equal sample size 5 in both groups. The relations of concentrations of two lipids in a pair can be expressed in a two-way ANOVA model:

$$z_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{8.1}$$

where z_{ijk} denotes the concentration of the i^{th} lipid in the k^{th} sample within the j^{th} treatment.

μ is the unknown true overall mean in the population.

α_i is the i^{th} role lipid effect (i.e. role is product or reactant).

β_j is the j^{th} (WT or MT) treatment effect.

γ_{ij} is the interaction between lipid roles and the treatment types.

The effects α_i, β_j and γ_{ij} are assumed to be fixed effects.

ε_{ijk} is a random component, assuming that $\varepsilon_{ijk} \xrightarrow{iid} N(0, \sigma^2)$ with mean 0 and variance σ^2.

The following steps illustrate the ANOVA approach for finding significant reactant and product pairs on the pathway.

1. All lipid pairs are screened by the relations according to y = 2. Those lipid pairs will be discarded if the screening scheme does not hold.

2. The hypothesis $H_0 : \mu_{11} = \mu_{12} = \mu_{21} = \mu_{22}$ is tested to detect the treatment effect from the model (8.1). If there is no significant differences in the four means in μ_{ij}, testing the interaction will be meaningless.
3. If the means are significantly different, the interaction is tested with the hypotheses H_{0i}: There is no significant interaction between the role of a lipid and the treatment type versus H_{Ai}: There is significant interaction between the role of a lipid and the treatment type. The interaction test in (8,1) can be expressed as H_{0i}: $\gamma_{ij} = 0$ versus H_{0i}: $\gamma_{ij} \neq 0$. The index *i* is indexing the various hypotheses being tested for different pairs of lipids.
4. A list of significant lipid pairs is given by using local *fdr* at three different levels, 0.001, 0.01 and 0.05.

In the following section, all 9 lipidomic experiment datasets are analyzed by using this ANOVA approach.

8.2. ANOVA Interaction Test Results for the 9 Lipid Datasets

Figure 8.1 shows a typical interaction plots from a lipid pair. The non-parallel or significant interaction for the lipid pair PE34_1 and PS34_2 with PE34_1 as the reactant and PS34_2 as the product with $y = 2$ in which the conditions $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ hold. The red line stands for the WT group, and the black line stands for the MT type group. The two points in A are WT mean $\bar{z}_{A1\bullet}$ (i.e., red triangle) and MT type mean $\bar{z}_{A2\bullet}$ (i.e., solid black dot), respectively. Similarly, the two points for B are WT mean $\bar{z}_{B1\bullet}$ (i.e., red triangle) and $\bar{z}_{B2\bullet}$ (i.e., solid black dot).

All 9 lipid experimental datasets are analyzed in the same manner. Table 8.2 shows the top 6 lipid pairs along with their tg, SSD, and p values from the interaction test that includes the corresponding *lfdr* values. Many lipid pairs show very small P values and the *lfdr* values. There are different numbers of significant lipid pairs in each of the 9 lipid datasets. The fad2 dataset has the most significant results.

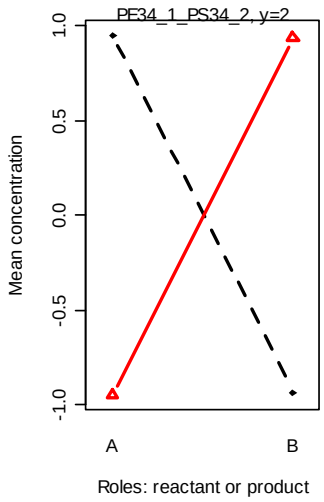

Figure 8.1: Interaction plots for a lipid pair with a significant interaction
The red line shows the WT group and the black dashed line shows the MT group. The significant interaction from the pair of A: PE34_1 and B: PS34_2 from the $y = 2$ pairs which satisfy s $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$.

Table 8.1: The top 6 significant lipid pairs from the results in *fad2* data with *lfdr* < 0.001
All the significant pairs are saved in different Excel files for *lfdr* < 0.05, *lfdr* < 0.01 and *lfdr* < 0.001. The list is given in a descending order of the *lfdr* values in the last column.

AB.name	y	tg	SSD	pvals	*lfdr*
PE34_1_PS34_2	2	0.988241	2.664968	2.32E-16	3.14E-13
PC34_1_PS34_2	2	0.989289	2.66354	4.22E-16	3.14E-13
PC36_2_PS34_2	2	0.989738	2.662928	5.38E-16	3.14E-13
PE36_2_PS34_2	2	0.990124	2.662404	6.59E-16	3.14E-13
PE34_1_PS40_2	2	0.985919	2.661875	8.05E-16	3.88E-13
PI34_1_PS34_2	2	0.991131	2.661039	1.09E-15	3.88E-13

Figure 8.2 is a comparison of the number of significant lipid pairs from all 9 datasets at three different *lfdr* levels: 0.05, 0.01 and 0.001. From this figure, one can see that those mutations (i.e., genotypes) appear to have the most significant effect on lipid reactant-product pairs in fad2 and sfd3, as detected by the ANOVA interaction test. Meanwhile, in fad4 and sfd1 the genotypes have the least influence.

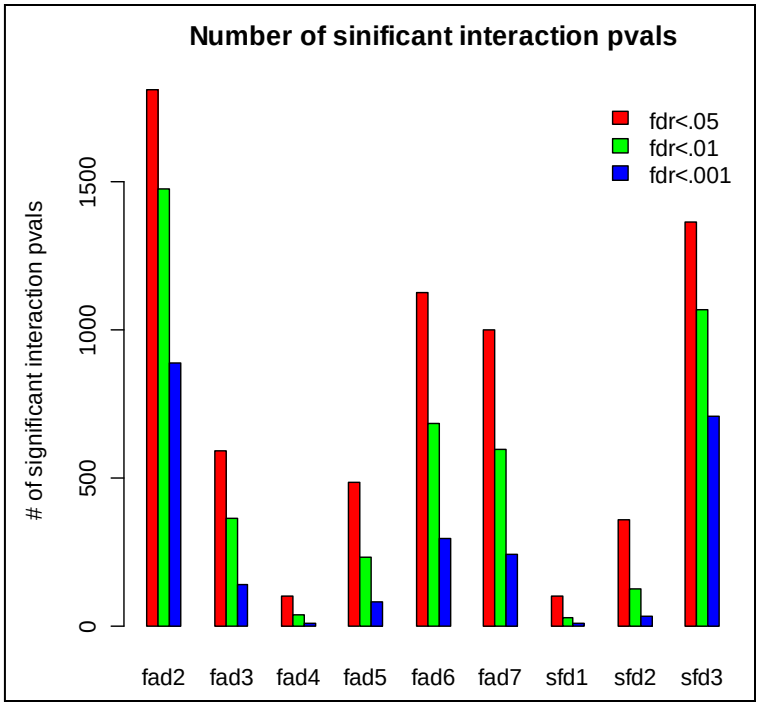

Figure 8.2: Comparison of the number of significant pairs in all 9 datasets
The 9 groups of bars show the number of significant lipid pairs in datasets fad2, fad3, fad4, fad5, fad6, fad7, sfd1, sfd2 and sfd3, respectively. In each group , the red bars are the number of significant lipid pairs for *lfdr* at level of 0.05, the green bars are the *lfdr* at level of 0.01, and the blue bars are the *lfdr* at level of 0.001.

8.3. ANOVA Interaction Test Results for the Roots-aerial data

The interaction test is also applied to the roots-aerial datasets from Fukushima et al. (2011). Treatment effects of *tt4* and *mtol* in the roots and aerial are compared with their WT (*Col0*) plants by using the interaction ANOVA model in (8.1). Figure 8.3 displays the number of significant metabolite pairs from the four datasets. The significant results are produced from the $y = 2$ screening, i.e., the condition that the mutation blocks a pathway. Figure 8.3 shows that the mutation has a larger effect in the roots than in the aerial portion at all the *lfdr* levels of 0.05, 0.01 and 0.001 by comparing the two left panels with the two right panels. In the roots part (two left panels), the two mutations *tt4* and *mtol* have similar effect with the WT. While in the aerial part (two right panels), the *mtol* effect is larger than that of *tt4*.

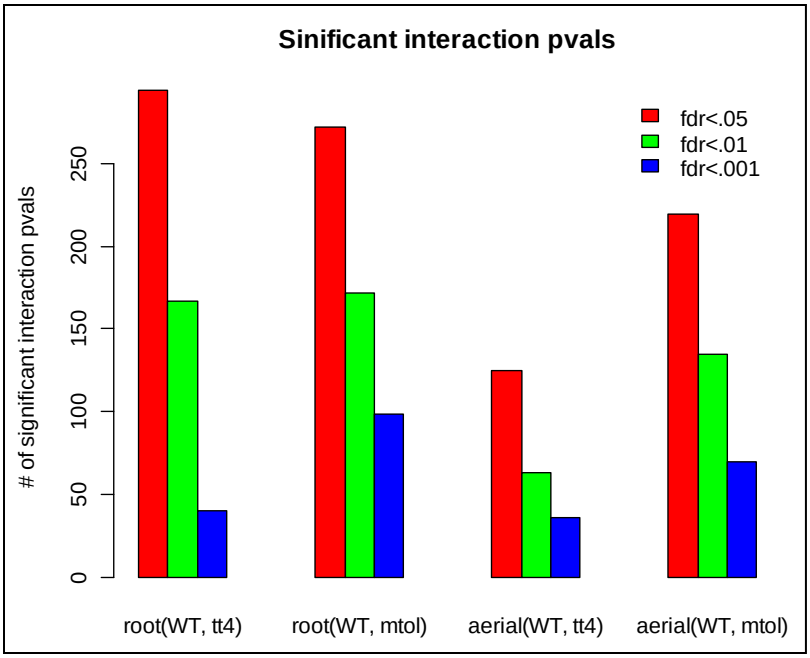

Figure 8.3: Comparison of significant metabolite pairs from the *roots-aerial* data using interaction p values

Four different datasets are shown in four panels: roots WT vs *tt4*, roots WT vs *mtol*, aerial WT vs *tt4*, and aerial WT vs *mtol*. In each dataset, the red bars stand for the frequency of the significant metabolite pairs by using *lfdr* at the level of 0.05, the green bars for the *lfdr* at the level of 0.01, and the blue bars for the *lfdr* at the level of 0.01.

This chapter has compared the mutation effects of the 9 lipid datasets by using the ANOVA approach in a two-way model including an interaction. In the previous chapters, several methods have been explored for the analysis of the lipidomic datasets in identifying the reactant and product pathways. In chapter 9, realistic lipidomics datasets will be simulated under the assumption that there is no mutation effect in each lipid pair. We will explore the characteristics of the simulated data to see how well the test statistics can reflect the relation between the means of a lipid pair. The Mixture Normal Bootstrap Null (MNBN) distribution will be used in the simulated data to evaluate the performance of this method.

Chapter 9 - Simulation of Realistic Data

9.1. Introduction

Many initial methods for analyzing gene-expression data made assumptions about independence among genes. It was shown that some methods still work well, on average, in the presence of dependence among genes or under assumptions of certain structures of dependence. It was later shown that some key results can be highly variable from study to study when gene expression data are dependent among genes. Recent work by Klebanov et al. (2007), Owen (2005), and Efron (2007; 2010) have investigated this dependence. Efron's work included an adjustment to estimated FDR values for dependence among genes. Paranagama (2011) explored a method to stabilize the variance of FDR estimates in the high dimensional data under dependencies. The metabolite pathway analysis is still a new research area in which there are not many rigorous statistical methods developed. In this chapter, characteristics of simulated data and performance of the method MNBN under a null hypothesis of equal means will be investigated in a simulation.

As methods for the analysis of gene expression data progressed over the past years, eventually attention turned to methods for simulating realistic high-dimensional data. Up to that point, data were simulated from multivariate normal distributions with restrictive dependence structures. Simulating more realistic gene expression data was considered in Gadbury et al., (2008). Simulating realistic lipidomic data is challenging. Still it will be necessary in order to evaluate the performance of new statistical methods for analyzing lipidomic data. Gadbury et al. (2008) proposed a plasmode method for generating data which were closer to the structure of real data. Paranagama (2011) extended this method and suggested a new plasmode method to simulate data with more original structure preserved in the datasets. Other techniques for simulating realistic data can be found in Hu et al. (2010) and Wang (2012).

9.2. Data Simulation Algorithms

To simulate realistic data for the lipidomics experiment, the two characteristics from the screening scheme in Figure 1.5 will be considered in the data simulation algorithm. The mean and standard deviation vectors from the real data will be utilized in the simulation of the realistic data for the purpose of preserving the real data structure. This will keep the mean changes in the

significant lipid pairs on the pathway to be consistent with the characteristics in the scheme. Two algorithms are considered.

Consider one of the datasets, fad4, as a multivariate data where there are 141 lipids and 5 samples in each WT and MT group. Let the mean and standard deviation vectors from the WT and MT groups be $\hat{\mu}_W$, $\hat{\mu}_M$, $\hat{\sigma}_W$, and $\hat{\sigma}_M$, respectively. Let the WT correlation matrix R_W be a 141 by 141 correlation matrix. The off-diagonal elements of the first 30 by 30 block diagonal matrix (in the red square) have correlation 0.5 in R_W as shown in (9.1). The other off-diagonal elements have zero values assuming independency in the lipid pairs.

$$R_W = \begin{bmatrix} 1 & 0.5 & ... & 0.5 & 0 & ... & 0 \\ 0.5 & 1 & ... & 0.5 & 0 & ... & 0 \\ .. & ... & ... & 0.5 & 0 & ... & 0 \\ 0.5 & 0.5 & 0.5 & 1 & 0 & ... & 0 \\ 0 & 0 & 0 & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... & ... & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad (9.1)$$

There are two ways to place the mutation effects into the simulated data. One way is to use the 7 biologically functional lipid pairs from the actual fad4 dataset as a criterion in the simulated dataset since their means and variances are preserved in the simulation. It is denoted by pairs.sim. The other way is to place the mutation effect on 7 arbitrary lipid pairs by changing the sign of the correlations. The 7 arbitrary lipid pairs are then called pairs.corr.

The correlation matrix in the MT group R_M in (9.2) has a similar structure with R_W, with a 30 by 30 block diagonal matrix (in the red square), but the correlations in the first 7 lipid pairs in the first row and first column of R_M (in the blue square) has changed to -0.5 in an attempt to simulate biologically functional lipid pairs.

$$R_M = \begin{bmatrix} 1 & -0.5 & \dots & -0.5 & 0.5 & \dots & 0.5 & 0 & \dots & 0 \\ -0.5 & 1 & \dots & 0.5 & 0.5 & \dots & 0.5 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -0.5 & 0.5 & \dots & 1 & 0.5 & \dots & 0.5 & 0 & \dots & 0 \\ 0.5 & 0.5 & \dots & 0.5 & 1 & \dots & 0.5 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0.5 & 0.5 & \dots & 0.5 & \dots & 0.5 & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & \dots & \dots & \dots & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 & \dots & 1 \end{bmatrix}, \quad (9.2)$$

The covariance matrices $\hat{\Sigma}_W$ and $\hat{\Sigma}_M$ for the WT and MT groups are computed by $\hat{\Sigma}_W = \hat{\sigma}_W \cdot R_W \cdot \hat{\sigma}_W^T$ and $\hat{\Sigma}_M = \hat{\sigma}_M \cdot R_M \cdot \hat{\sigma}_M^T$. Then, the simulated fad4 data are produced by using a multivariate normal distribution with the following forms:

$$X_W \sim N_{141}(\hat{\mu}_W, \hat{\Sigma}_W) \text{ and } X_M \sim N_{141}(\hat{\mu}_M, \hat{\Sigma}_M), \quad (9.3)$$

The simulated "fad4 data" are formed with X_W as its WT group data and X_M as its MT group data.

Algorithm 1: Simulate realistic fad 4 data and get the distribution of test statistic R_T.

1. Compute the mean and standard deviations $\hat{\mu}_W$, $\hat{\sigma}_W$, $\hat{\mu}_M$ and $\hat{\sigma}_M$ from fad4 data in the WT and MT groups, respectively.
2. Simulate the correlation matrix R_W and R_M following the structure described above.
3. Compute the covariance matrices $\hat{\Sigma}_W$ and $\hat{\Sigma}_M$ in the WT and MT groups, respectively.
4. The simulated realistic data are produced from multivariate normal distributions in (9.3) and the dataset is denoted by sim.real.
5. Center and scale the dataset sim.real using $z_{ij} = \frac{x_{ij} - \bar{x}_{\bullet\bullet}}{s}$ to get the mean 0 and standard deviation 1 for each lipid species. x_{ij} and z_{ij} denote the lipid before and after scaling, respectively.
6. Proceed to create the distribution of R_T as described in earlier chapters.

In Algorithm 1, the simulated realistic data are very close to the real data since it has preserved the mean and also the variance structure for each lipid. The simulated biologically biologically functional lipid pairs are generated by changing the sign of the correlation in the MT group to -0.5 for 7 lipid pairs to check whether correlation can be used as a criterion to find the significant lipid pairs on the lipidomics pathway. If the 7 arbitrary biologically functional lipid pairs (pairs.corr) are at the top of the list of results, we may conclude that altering correlations is a useful method to create lipid pairs in simulated data. Since the 7 simulated biologically functional lipid pairs (pairs.sim) have the same mean and variances in the simulated data, the 7 simulated biologically functional lipid pairs with the 7 arbitrary biologically functional lipid pairs will be used as a criterion to evaluate the method and to investigate the properties of the simulated datasets.

The simulated dataset (sim.real) from Algorithm1 will be used as "real data" in the simulation. Then, how to simulate a null dataset under the null hypothesis of no mutation effect in each lipid pair? The following Algorithm 2 proposes a method to generate a null dataset from which the bootstrap samples are produced.

Algorithm 2: Simulate a null dataset and produce the bootstrap null distribution for R_T under the null hypothesis $\mu_F = \mu_G$.

1. Compute the covariance matrices $\hat{\Sigma}_W$ and $\hat{\Sigma}_M$ using the following formulas $\hat{\Sigma}_W = \hat{\sigma}_W \cdot R_W \cdot \hat{\sigma}_W^T$ and $\hat{\Sigma}_M = \hat{\sigma}_M \cdot R_W \cdot \hat{\sigma}_M^T$. Common correlation matrix R_W is used in both covariance matrices.
2. The simulated null distribution data are produced from (9.3) and the dataset is denoted by sim.null.
3. Let the sim.null be a 141 by 10 matrix with the 141 rows representing the 141 lipid species and the 10 columns representing 5 WT samples and 5 MT samples. The data sim.null will be a template from which the 200 bootstrap null samples are generated.
4. Let the WT sample data in the sim.null dataset be $w_1, w_2, ..., w_n$ with a population mean μ_F and a sample mean $\overline{w}$. Let the MT data be $m_1, m_2, ..., m_n$ with a population mean μ_G and a sample mean $\overline{m}$.

5. Transform the data by using $\tilde{w}_i = w_i - \overline{w} + \bar{x}$ and $\tilde{m}_i = m_i - \overline{m} + \bar{x}$, $i = 1,2,...,n$, where $\bar{x}$ is the mean of the combined sample.
6. Let B = 200 and $b = 1, 2,..., B$ denote the b[th] bootstrap sample ($\tilde{w}^{*b}, \tilde{m}^{*b}$), where $\tilde{w}^{*b}$ is sampled with replacement from $\tilde{w}_1,...,\tilde{w}_5$ and $\tilde{m}^{*b}$ is sampled with replacement from $\tilde{m}_1,...,\tilde{m}_5$.
7. Center and scale samples $\mathrm{x}^{*b} = (\tilde{w}^{*b}, \tilde{m}^{*b})$ by using $z_{ij}^* = \frac{x_{ij}^* - \bar{x}_{\bullet\bullet}^*}{s^*}$ to get the mean 0 and standard deviation 1 for each lipid species. x_{ij}^* and z_{ij}^* denote the bootstrap data before and after scaling, respectively.
8. Pair every reactant A^* with every product B^* to get approximately $2\binom{141}{2} = 19740$ lipid pairs $\mathrm{A}^* \rightarrow \mathrm{B}^*$.
9. Screen the relationships of $\bar{z}_{A1\bullet}^* < \bar{z}_{A2\bullet}^*$ and $\bar{z}_{B1\bullet}^* > \bar{z}_{B2\bullet}^*$ according to the value $y^* = 2$ for any arbitrary lipid pairs. In total, about K = 4600 lipid pairs satisfy the conditions of $y^* = 2$.
10. Calculate the statistic R_T^* for each lipid pair $\mathrm{A}^*\mathrm{B}^*$ from the scaled bootstrap sample z^{*b} with y^* = 2. The test statistic is a vector with about 4600 elements from each of the b[th] bootstrap samples, i.e., $\mathrm{R}_\mathrm{T}^{*\mathrm{b}} = (\mathrm{R}_\mathrm{T}^{*1}, \mathrm{R}_\mathrm{T}^{*2},..., \mathrm{R}_\mathrm{T}^{*4600})$. The R^* statistic is defined as

$$\mathrm{R}^* = (tg^* - 1)^2 + (SSD^* - 2.684))^2.$$

11. Repeat steps 6 through 11 200 times to get about 4600×200 bootstrap statistics to determine the null distribution for the statistic R_T^*.

The reason for using the common correlation matrix in step 1 in both WT and MT is that in the screening scheme, if there is no mutation effect in the AB lipid pair and the mutation did not block the pathway in the MT group, then $\mathrm{A_m} \rightarrow \mathrm{B_m}$ will proceed in the same rate as $\mathrm{A_w} \rightarrow \mathrm{B_w}$. So a common correlation matrix is used when producing null data for WT and MT. Note that to create the null data that satisfy the scheme under no mutation effect, the mean structure must be removed from the biologically functional lipid pairs, which is implemented in step 9 of Algorithm2.

9.3. Simulation Using fad4 Dataset

The simulation is done with sample size n = 5 and n = 20. To investigate the correlation effect on the simulated datasets, two forms of the correlation matrices R_W have been used. One is in the form of (9.1) with a 30×30 block values of 0.5. Another form is similar with the form of (9.1), but the 30×30 block of values change to 0.8. The correlation matrix in the MT group is similar with their corresponding WT correlation matrix R_W except that the 7 values for the arbitrary biologically functional lipid pairs (pairs.corr) change to negative.

Simulation 1: Sample size n = 5 with correlation values 0.5 in the block matrix in R_W.

1) Choose WT correlation matrices R_W (9.1) and R_M (9.2).
2) Compute $\hat{\Sigma}_W = \hat{\sigma}_W \cdot R_W \cdot \hat{\sigma}_W^T$ and $\hat{\Sigma}_M = \hat{\sigma}_M \cdot R_M \cdot \hat{\sigma}_M^T$ to get the simulated real dataset sim.real by following Algorithm 1.
3) Compute $\hat{\Sigma}_W = \hat{\sigma}_W \cdot R_W \cdot \hat{\sigma}_W^T$ and $\hat{\Sigma}_M = \hat{\sigma}_M \cdot R_W \cdot \hat{\sigma}_M^T$ to get the simulated null dataset sim.null by following Algorithm 2, and also to get the bootstrap distribution for R_T.
4) Overlay the R_T distribution from sim.real with the bootstrap distribution. Fit the bootstrap distribution using two-component mixture model.
5) Get the P values from the two-component mixture model. Control the false discovery rate by local *fdr* at the level of 0.05 to find a list of findings.

Simulation 2: Sample size n = 5 with correlation values 0.8 in the block matrix R_W.

Simulation 2 is similar with simulation 1 with sample size n = 5 except that the off-diagonal values in the correlation matrices change to 0.8 in both R_W (9.1) and R_M (9.2).

Simulation 3: Sample size n = 20 with correlation values 0.5 in the block matrix R_W.

Simulation 3 is similar with simulation 1 with the same correlation structure in both WT and MT groups except that the sample size increases to n = 20.

Simulation 4: Sample size n = 20 with correlation values 0.8 in the block matrix R_W.

Simulation 4 is similar with simulation 2 with the same correlation structure in both WT and MT groups except the sample size increases to n = 20.

9.3.1. The Characteristics of the Simulated Datasets

Figure 9.1(a) shows the scatter plot of tg ratio versus SSD in the actual data from fad4. The red points on the peak are the biologically functional lipid pairs (tg, SSD). The red data points show a region where the significant findings should be located if the lipid pairs satisfy the screening scheme. Also, for the red points their SSD values are close to max(SSD), i.e., 2.684, when sample size is 5, and their tg values are close to 1. All these properties are reflected in the R statistics formula which was defined in (3.6). In Figure 9.1(b), the R statistics are transformed to R_T statistics. The red lines which represent the biological biologically functional lipid pairs R_T statistics indicate that the "interesting" lipid pairs are shown by the larger values of R_T.

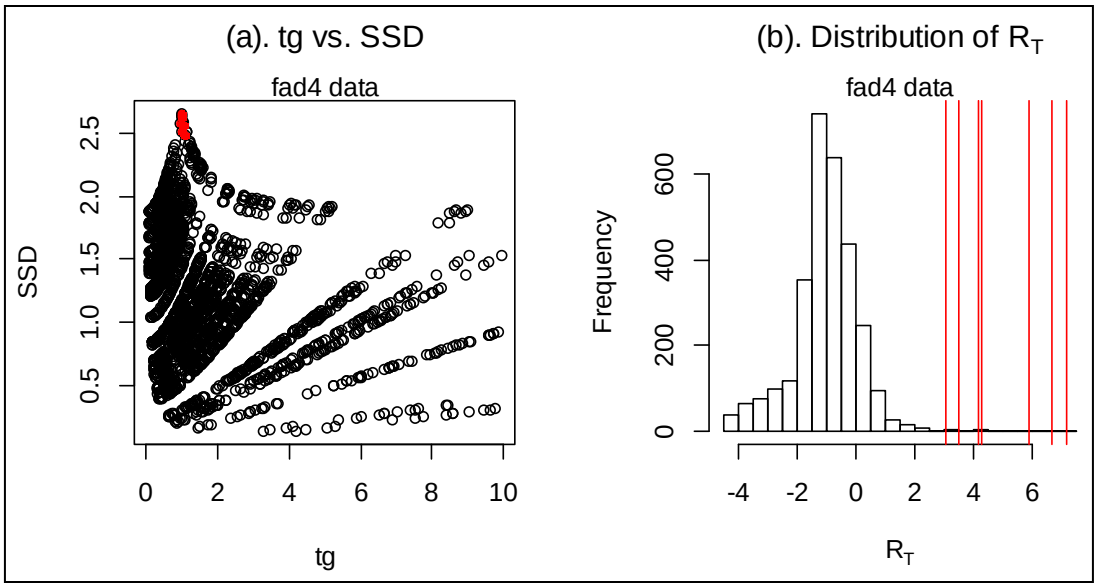

Figure 9.1: The scatter plot tg versus SSD and the R_T distribution in the actual dataset fad4
(a). Scatter plot of tg vs. SSD in fad4. The red data points are the tg vs. SSD from the biologically functional lipid pairs. (b). The distribution of R_T. The red lines represent the biologically functional lipid pairs' R_T statistics.

Figure 9.2 shows the scatter plots of tg versus SSD from the sim.real data (in Algorithm 1) in each simulation. We can see that the simulated biologically functional lipid pairs (pairs.sim in red points) appear on the top of each scatter plot. The positions of the simulated biologically functional lipid pairs (pairs.sim) are similar with those in the actual fad4 data scatter plot shown in Figure 9.1(a). That means the biologically functional lipid pairs are a good indication of the lipid pairs on the pathway in both simulated data and real data.

The arbitrarily paired lipid pairs (pairs.corr, orange points) did not appear in simulation 2 because those pairs are screened out by the $y = 2$ criterion according to the mean relations in the

screening scheme. In all other plots in Figure 9.2, the arbitrary biologically functional lipid pairs, pairs.corr, are randomly scattered without concentrating on a specific position. The appearance of the pairs.corr (orange points) did not reflect the characteristics of the actual data like the simulated biologically functional lipid pairs (red points). Hence, the simulated biologically functional lipid pairs (pairs.sim) will catch the mutation effect more accurately than the arbitrary pairs (pairs.corr).

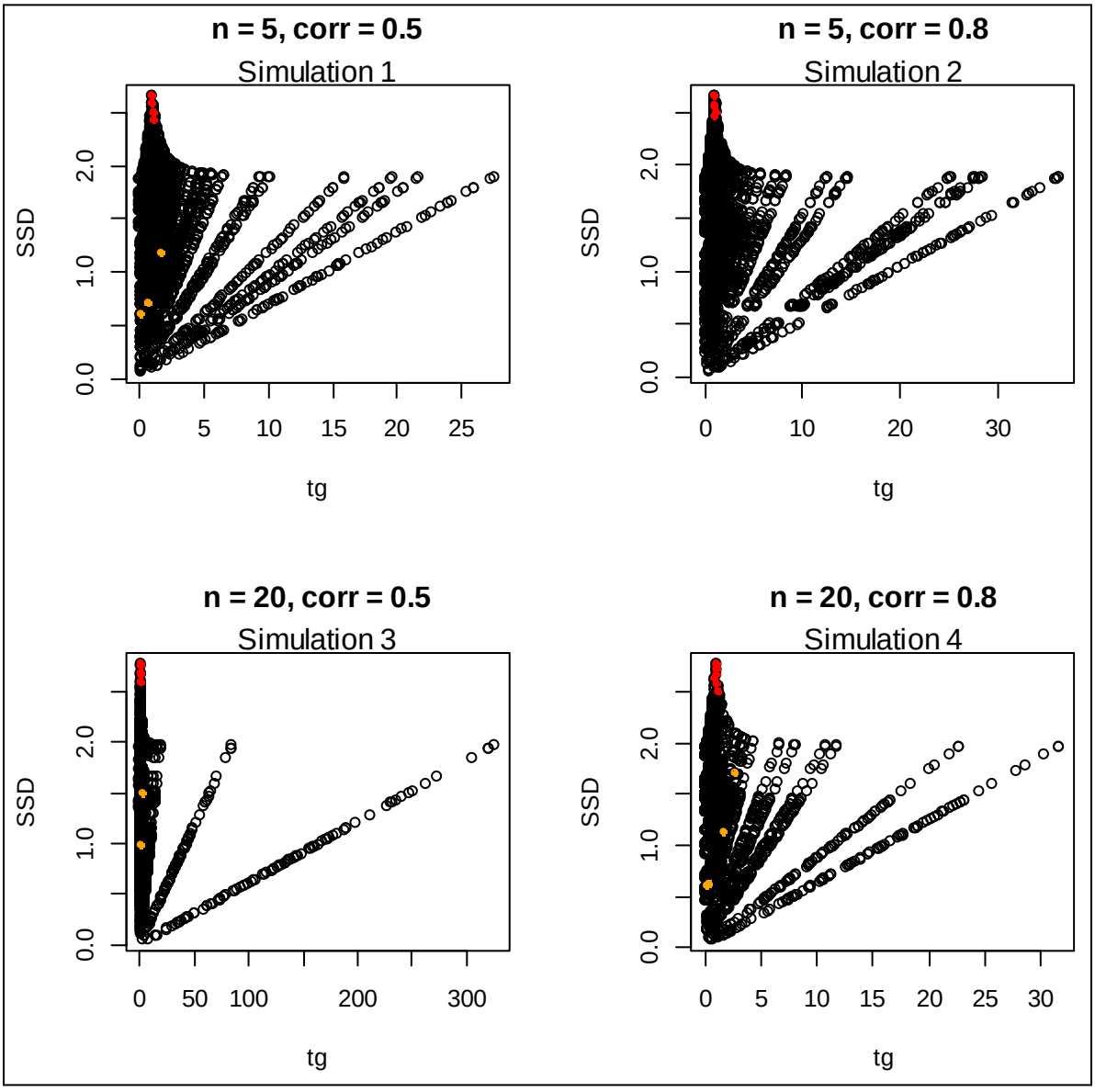

Figure 9.2: The scatter plot of tg versus SSD from the four simulations
The scatter plots in the four panels are all produced from the simulated real data, i.e. sim.real, from the four simulations. In each panel, the red points show the pairs from the pairs.sim. The orange points stand for the arbitrary biologically pairs from pairs.corr.

Now, let us compare the R_T distribution in the actual data in Figure 9.1(b) and in the four simulations in Figure 9.3. In Figure 9.3, the R_T values for the pairs.sim lipid pairs (red line) have resemble the pattern in Figure 9.1(b) in modeling large values of the statistics.

The pair.real.corr arbitrary lipid pair statistics R_T in each panel cannot model large R_T statistics. Also, in simulation 2 the arbitrary biologically functional lipid pairs are screened out by the y = 2 criterion. Therefore, the arbitrary biologically functional lipid pairs cannot reflect the characteristics of the datasets. Furthermore, the correlation does not show a strong effect on the lipid pairs. Extensive simulation study should be conducted to investigate the correlation effect on the pathway to get a final conclusion.

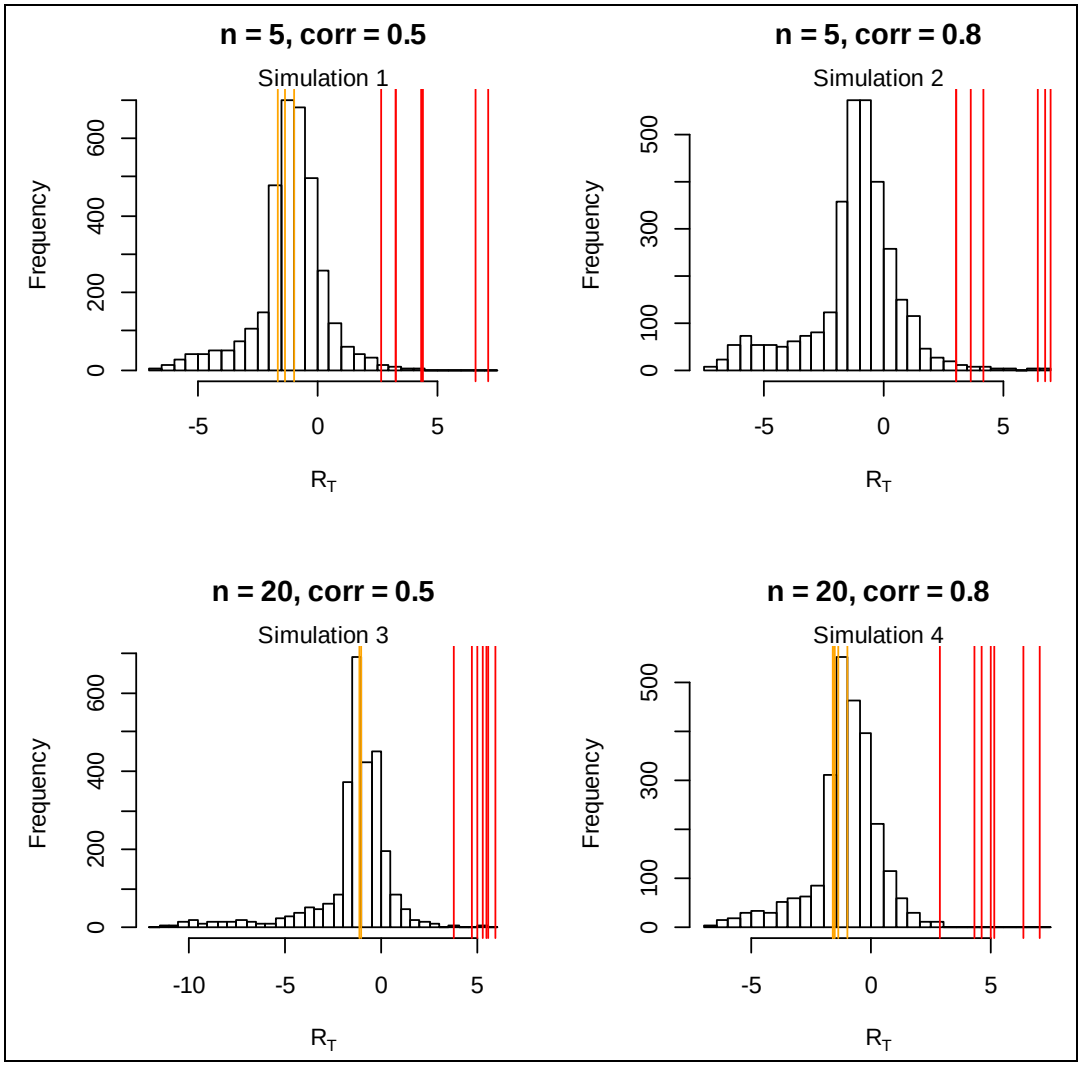

Figure 9.3: The R_T distribution in the four simulations
The red lines show the R_T statistics from the simulated biologically functional lipid pairs, pairs.sim. The orange lines represent the R_T statistics from the arbitrary biologically functional lipid pairs from pairs.corr.

This simulation method provides a useful tool to simulate realistic data that is close to the actual fad4 data. The characteristics are similar in both real data and simulated datasets. In the next section, the MNBN method is applied to all four simulations.

9.3.2. The MNBN Method to Fit to the Bootstrap Distributions

Figure 9.4 shows the normal mixture model fitting to the actual fad4 dataset using two-component mixture model. The green histogram is the R_T distribution from fad4. The orange histogram shows the bootstrap samples distribution. Three- and four- component mixture models are also explored in the MNBN method. But two-component mixture normal model has been used to fit the bootstrap null distribution in the MNBN method since two-component mixture model can capture the shape of the empirical bootstrap distribution.

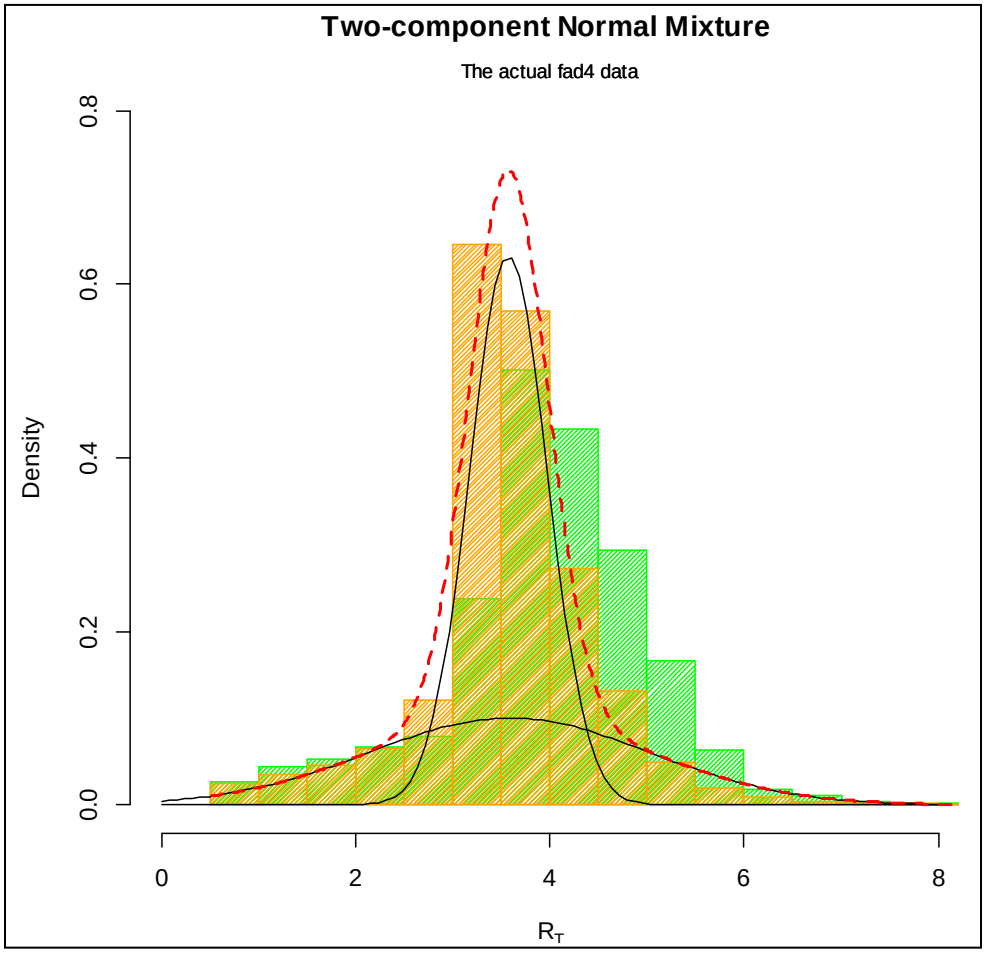

Figure 9.4: Normal mixture model fit to the actual fad4 dataset
The black curves are the normal component densities and the red dashed lines are the mixture model density. The green histogram is the R_T distribution in the actual data fad4. The orange histogram is the empirical bootstrap distribution.

Figure 9.5 shows the two-component normal mixtures fit to the empirical bootstrap distributions in the four simulations using the MNBN method. The results produced from those four two-component mixture models will be discussed in the next section.

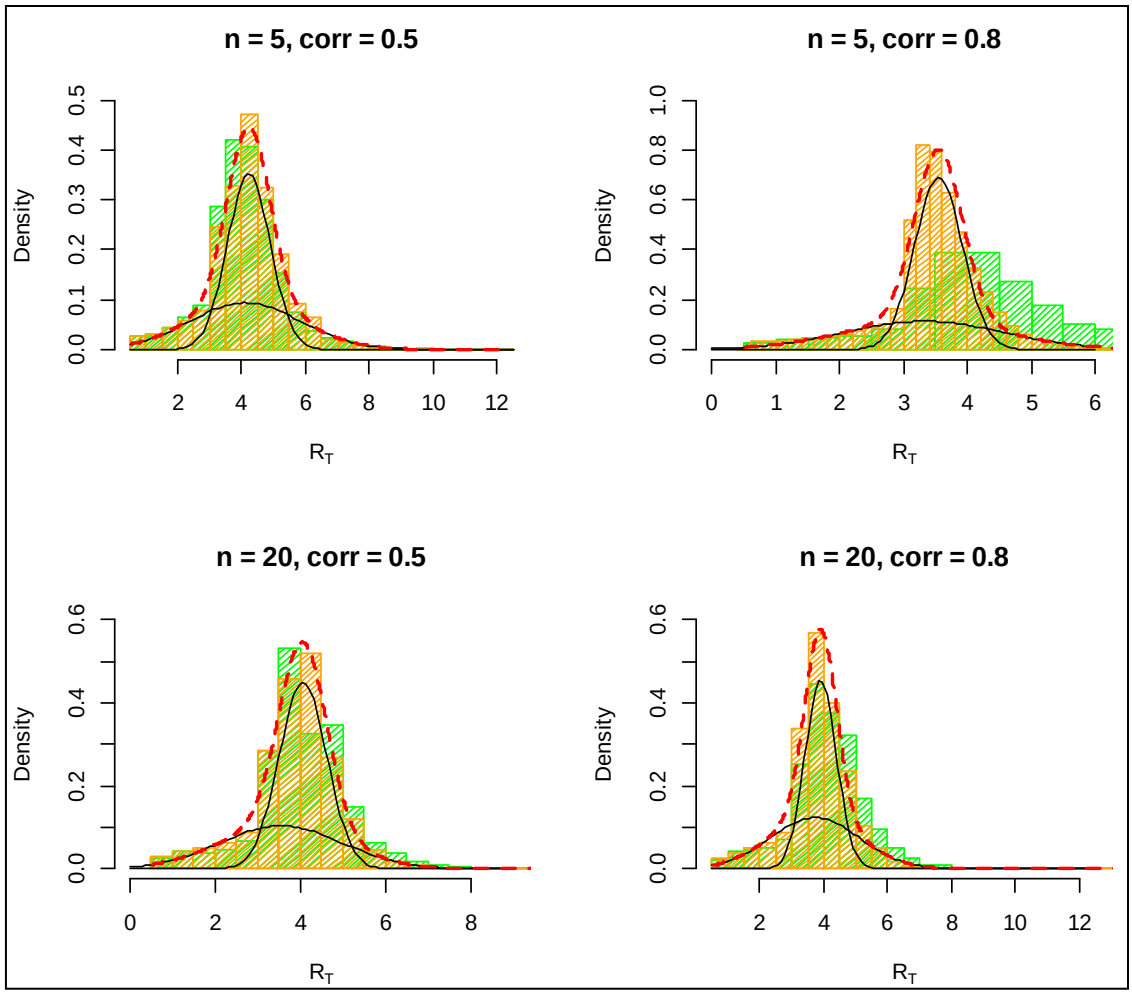

Figure 9.5: Two-component normal mixture model fit in the four simulations
In each panel, the green histograms stand for the distribution of R_T in the simulated real data from sim.real. The orange histograms represent the bootstrap samples produced from the bootstrap. The two black curves are the two normal densities and the red dashed curves are the density of the two-component mixture models.

9.4. Results and Discussion

The fitting results of the two components with the MNBN method are listed in Table 9.1 after local *fdr* multiple adjustments at level of 0.05.

In Table 9.1, column 1 shows the four simulations with different sample sizes and correlation structures. Column 2 is the number of lipid pairs in the simulated real data set (sim.real) which will be used as the "real" data to overlay with the bootstrap distribution using

the MNBN method. Column 3 shows the total number of findings produced from the MNBN. Column 4 shows the number of simulated biologically functional lipid pairs (pairs.sim) appeared in the final findings. The last column shows the number of the arbitrary biologically functional lipid pairs (pairs.corr) in the final findings from column 3.

In the last column of Table 9.1, the arbitrary biologically functional lipid pairs (pairs.corr) do not appear in any of the results. The correlations for the 7 simulated arbitrary pairs were changed from 0.5 to 0.8. The correlation change in the simulated biologically functional lipid pairs does not promote the arbitrary pairs to the top of the list findings. In all four simulations, the correlation does not seem to play a vital role in determining the reactant-product lipid pairs.

Table 9.1: The results from the 4 simulations using the MNBN method
The first column shows the four simulations. The second column represents the total number of lipid pairs in the simulated real data (sim.real). Column 3 shows the total number of lipid pairs in the final results. Column 4 shows the number of simulated biologically functional lipid pairs (pairs.sim) that appeared in the final results. The last column is the number of the arbitrary biologically functional lipid pairs (pairs.corr) in the final results.

Simulation	# of lipid pairs in sim.real	Total number of significant results	# of pairs.sim	# of pairs.corr
n = 5, corr = 0.5	3323	0	0	0
n = 5, corr = 0.8	2935	0	0	0
n = 20, corr = 0.5	2606	10	7	0
n = 20, corr = 0.8	2468	8	6	0

Some summary remarks

The simulation algorithms in this chapter provide a tool to simulate realistic lipidomic pathway data. The simulated realistic data captured the characteristics of the real data. The simulated null datasets are produced in the situation when the null hypothesis, H_0: no mutation effect, is true. The method MNBN is more precise (in capturing the biologically functional lipid pairs) with an increased sample size. Since the results in Table 9.1 are from four simulations, a concrete conclusion should be drawn from extensive simulations. Fukushima et al. (2011) addressed that strong correlation between the metabolite pairs is an indication of the pathway. Raamsdonk et al. (2001) investigated the metabolite pathway using the concentration change in WT and MT groups as introduced in chapter 2. In this simulation study, it shows that the

concentration change in the WT and the MT groups seems more relevant to the pathway than correlation analysis.

Chapter 10 - Summary and Future Work

10.1 Summary of This Dissertation

Compared to previous methods for find a metabolic pathway, the method used in this research not only focuses on the mean concentration changes in a lipid pair, but also determines the reaction direction from a reactant (A) to a product (B), making use of the mutation effect in blocking a reaction. The main idea, which is reflected from the screening scheme, has been studied in an exploratory data analysis in chapter 3. Some numerical facts and relations between the means for a lipid pair were found. Three summary statistics tg, SSD and R_T were established and four methods proposed in chapters 4, 5, 6, 7 and 8. Chapter 4 explored a bootstrap procedure to find the parametric null distribution of the three test statistics. In this bootstrap method, a criterion, Minimum Kolmogorov-Smirnov (K-S) D statistics, was introduced and utilized to select a candidate parametric model as the null distribution of the test statistic. The parametric bootstrap null distributions from chapter 4 presented some challenges in capturing the structure in the data. This lead to the investigation of a mixture model fit to the bootstrap null distribution in chapter 5, where a two-component mixture normal model has used to fit the bootstrap null distribution under the restricted null hypothesis of $F = G$. Chapter 5 also presented the randomization test for testing the treatment effect. Two datasets, fad2 and fad4, showed the strongest and weakest treatment effect among all datasets studied in this research. The data fad2 has the largest number of lipid pairs in the final findings as expected. When a two treatment comparison test between WT and MT is performed on the 141 lipids, it can be seen that most are significantly affected by the mutation in fad2 data.

For the application herein and considering the challenge of producing a valid null distribution, an alternative approach has been considered in chapter 6 under a different null hypothesis, that is, one of equal means $\mu_F = \mu_G$ rather than equal distribution as was done in chapters 4 and 5. The bootstrap methods, using the parametric distribution to fit the bootstrap samples as shown in chapter 4 and using the mixture normal distribution to fit the bootstrap null distribution as shown in chapter 5, are compared under the equal mean null hypothesis.

To fit the statistic R_T to the mixture model, an analogous approach to Efron's "*empirical null distribution*" has been introduced in chapter 7, using a normal mixture model with two-, three-, and four-component. The number of components in a mixture model is tested based on a

parametric bootstrap procedure. A three-component mixture normal is selected by the bootstrap method and used to fit the data. In this model, one component models the bigger values of R_T for which the null hypothesis is false, and the other two components model results for which it is true. Some results have been shown by using the posterior probability that a lipid pair is affected by the mutation at a particular R_T. In chapter 8, a more conventional approach has been explored by using a two-way ANOVA model with an interaction term. The interaction reflects the change in lipid concentrations that are of interest in this research. The results produced from this method after local *fdr* adjustments are similar with those from the other three methods.

As methods for the analysis of gene expression data progressed over the past years, eventually attention turned to methods for simulating realistic high-dimensional data. Simulations are performed in chapter 9 to illustrate the method using mixture normal to fit the bootstrap null distribution under the equal mean assumption.

10.2 Future Direction

10.2.1. Intersection-Union Test

In chapter 4, a more restrictive null hypothesis $F = G$ was applied to a specific null hypothesis for each lipid in the data sets. The problem is that the null distribution of the test statistics R_T may deviate quite substantially from the actual distribution of R_T seen in data. The results suggest strong mutation effects in the fad2 dataset. A less restrictive null hypothesis would be,

$$\begin{aligned} H_0 &: \mu_{Aw} \geq \mu_{Am} \text{ or } \mu_{Bw} \leq \mu_{Bm} \\ H_A &: \mu_{Aw} < \mu_{Am} \text{ and } \mu_{Bw} > \mu_{Bm}. \end{aligned} \tag{10.1}$$

The above test was first defined in the Intersection-Union Tests (Berger 1997; Berger and Hsu 1996) which rejects the H_0 only if all the tests are rejected (http://statweb.calpoly.edu/jdoi/web /research/iut_paper_proc.pdf).

Since we are interested in whether A and B are a reactant and product pair candidate or not, the alternative hypothesis in (10.1) reflects the screening scheme. Compared to the null hypothesis in (4.1), this hypothesis is less restrictive on the relations of the means for each lipid in a sense that only one of the two events, $\mu_{Aw} \geq \mu_{Am}$ or $\mu_{Bw} \leq \mu_{Bm}$, is required to be true, leading to less restrictive conditions on the four means in the lipid pair. The appropriate

bootstrap procedure would need to be explored that would bootstrap a null distribution for this set of hypotheses, and the characteristics of the test investigated for its use on high-dimensional data.

10.2.2. Dependence in the Data

An independence assumption was made in fitting the mixture model to data from lipid pairs. This was needed to write out the likelihood of the data. This likelihood expression is not technically the right likelihood, but it is considered a measure of relative model fit as it has been in other work on gene expression experiments. The lipidomic data described here are correlated in two ways. The first way was exploited when using the reactant-product pathways to identify interesting results. However, the fact is that many lipids may be on the same pathway and, therefore, there is likely to be correlation among pairs of lipids. Also, a common lipid may be present in many pairs giving another degree of dependence among data from paired lipids. What role does dependence play in characterizing reactant-product pairs or to what extent are results affected? Is an adjustment needed or will a new method be required to consider lipid reactions in longer chains? These questions are likely to be difficult to answer. In the simulation study in chapter 9, different correlation structures were applied to simulate realistic data to explore characteristic of the datasets. More extensive simulation should be done to investigate dependency in the data.

10.2.3. Variance Structure in the Lipid Pairs

The relations between the four means of a lipid pair is reflected in the screening scheme and also used in all the methods to develop the statistics in order to find the significant findings in the pathway. One may wondering if changes in variance may also help to identify lipid pairs whose pathway is affected by the mutation.

A initial exploration is done to evaluate equal variances in the WT and MT group. The F test statistic which is a ratio of the variance of the WT group σ_w^2 to the variance of the MT group σ_m^2 is used as a test statistic. The distributions of the test statistics F and the p value distributions are shown in Figure 10.1.

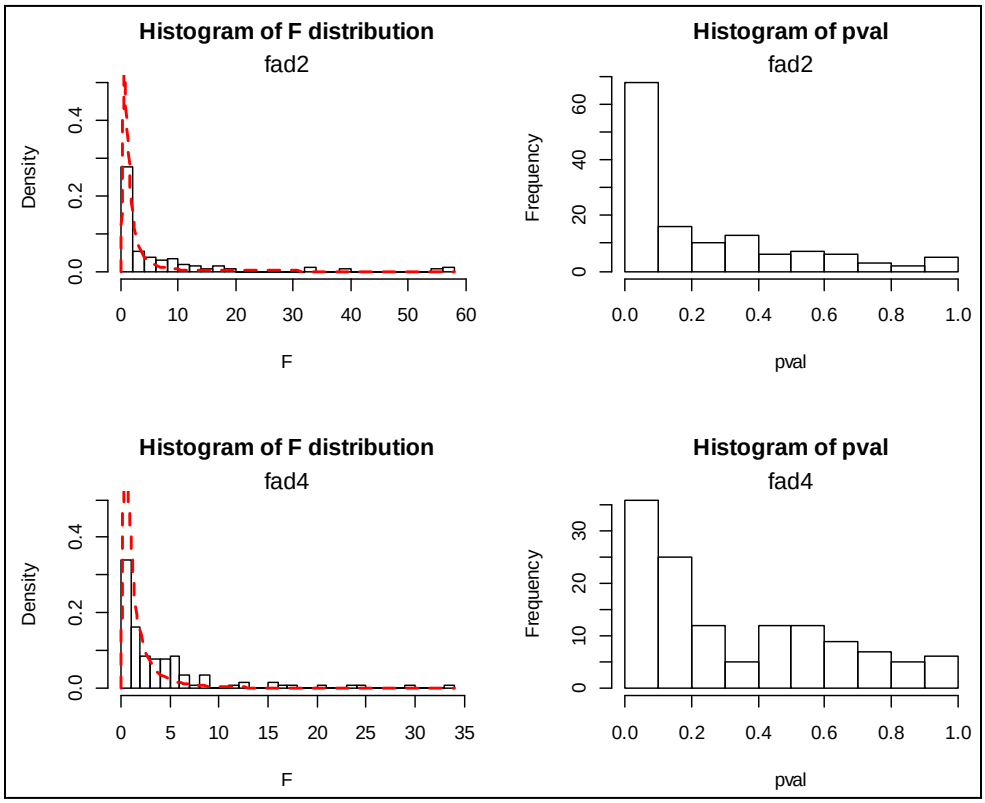

Figure 10.1: The distribution of statistic F from the equal variance test
The left panels show the distributions of the F test statistics (the histogram) from fad2 (top) and fad4 (bottom). The red dashed lines show the theoretical F distribution with *df1* = 4 and *df2* = 4. The right panels show the p value distributions from the equal variance test in fad2 (top) and fad4 (bottom).

The graph show some evidence of unequal variances across MT and WT groups for some lipids. To what extent, this manifests itself in lipid pathways that are modified by a mutation is a subject of future work.

References

Alberts B, Johnson A, Lewis J, Raff, M., Roberts, K., and Walter P. (2002). Molecular Biology of the Cell. *4th edition,* New York: Garland Science.

Allen, E., Moing, A., Ebbels, MD T., Maucourt, M., Tomos, A D., Rolin, D., and Hooks, M.A. (2010). Correlation Network Analysis reveals a sequential reorganization of metabolic and transcriptional states during germination and gene-metabolite relationships in developing seedlings of Arabidopsis. *BMC Systems Biology.* 4, 62.

Allen, J., Davey, H. M., Broadhurst, D., Heald, J. K, Rowland, J. J., Oliver, S. G. and Kell, D. B. (2003). High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nature Biotechnology,* 21*,* 692- 696.

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A., et al. (2002). A Mixture Model Approach for the Analysis of Microarray Gene Expression Data. *Computational Statistics and Data Analysis*. 39, 1-20.

Arondel, V., Lemieux, B., Hwang, I., Gibson, S., Goodman, H.M. and Somerville, C.R. (1992), Map-based cloning of a gene controlling omega-3 fatty acid desaturation in *Arabidopsis*. *Science*, 258, 1353-1355.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57, 289-300.

Berger, R.L. (1997). Likelihood Ratio Tests and Intersection-Union Tests. *Advances in Statistical Decision Theory and Applications*, Boston: Birkhauser. pp. 225 - 237.

Berger, R.L. and Hsu, J.C. (1996). Bioequivalence trials, intersection- union tests, and equivalence confidence sets. *Statistical Science*, 11, 283-319.

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J, Pedro Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey R. N., Lange, B. M., Wurtele, E. S. and Sumner, L. W. (2004). Potential of metabolomics as a functional genomics tool. *TRENDS in Plant Science,* 9, 418-425.

Blei, I., Odian G. (2006). General, organic, and biochemistry. Second edition, New York: W.H. Freeman and Company.

Broberg, P. (2005). A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*,6:199.

Brügger, B., Erben, G., Sandhoff, R., Wieland, F.T., Lehmann W.D. (1997). Quantitative analysis of biological membrane lipids at the low picomole level by nano-electrospray ionization tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 94, 2339-2344.

Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee1, S.Y., Tissier, C., Zhang, P. and Karp, P.D. (2006). MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*. 34, D511-D516.

Devaiah, S. P., Roth, M. R., Baughman, E., Li, M., Tamura, P., Jeannotte, R., Welti, R.,and Wang, X. (2006). Quantitative profiling of polar glycerolipid species and the role of phospholipase Dα1 in defining the lipid species in Arabidopsis tissues. *Phytochemistry*, 67, 1907-1924.

Dixon, R. A., Gang, D. R., Charlton, A. J., Fiehn, O., Kuiper, H. A., Reynolds, T. L., Tjeerdema, R. S., Jeffery, E. H., German, J. B., Ridley, W. P. and Seiber, J. N. (2006). Applications of Metabolomics in Agriculture. *Agricultural and Food Chemisty*, 54, 8984 - 8994.

Dunn, W. B., Ellis, D. I. (2005). Metabolomics: Current analytical platforms and methodologies. Trend in analytical chemistry, 24, 285 - 294.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*. 99, 96-104.

Efron, B. (2007). Correlation and large-scale simultaneous significance tesing. *Journal of the American Statistical Association*.102, 93-103.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates (with discussion and Rejoinder). *JASA*, 1042-1069.

Efron, B., Tibshirani, R. (1993). An Introduction to the Bootstrap. Chapman & Hall/CRC.

Efron, B., Tibshirani, R., Storey, J.D, Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*. 96, 1151-1160.

Everitt, B. S. and D. J. Hand, (1981), Finite Mixture Distributions, Chapman and Hall.

Falcone, D.L., Gibson, S., Lemieux, B. and Somerville, C. (1994). Identification of a gene that complements an Arabidopsis mutant deficient in chloroplast omega 6 desaturase activity. *Plant Physiology*, 106, 1453-1459.

Fan, L. (2010). An exploratory method for identifying reactant-product lipid pairs from lipidomic profiles of wild-type and mutant leaves of *Arabidopsis thaliana*. Master report.

Fiehn, O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology,* 48, 155–171

Fiehn, O. (2006) Metabolite Profiling in Arabidopsis. *Arabidopsis Protocols 2nd edition. Methods in Molecular Biology series*, Humana Press, Totowa NJ, 439-447.

Forth, T., McConkey, G.A., Westhead, D.R. (2010). MetNetMaker: a free and open-source tool for the creation of novel metabolic networks in SBML format. *Bioinformatics*. 26, 2352-2353.

Fukushima, A., Kusano, M., Redestig H., Arita, M., Saito, K. (2011). Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. *BMC Systems Biology*, 5, 1-12.

Gadbury, G. L., Xiang, Q., Yang, L., Barnes, S., Page, G. P., Allison, D. B. (2008). Evaluating statistical methods using plasmode datasets in the age of massive public databases: An illustration using False Discovery Rates. *PLos Genetics*.

Gao, J., Ajjawi, I., Manoli, A., Sawin, A., Xu, C., Froehlich, J. E., Last, R. L. Benning, C. (2009). FATTY ACID DESATURASE4 of Arabidopsis encodes a protein distinct from characterized fatty acid desaturases. *The Plant Journal*, 60, 832–839.

Gibson, S., Arondel, V., Iba, K. and Somerville, C. (1994). Cloning of a temperature-regulated gene encoding a chloroplast omega-3 desaturase from Arabidopsis thaliana. *Plant Physiol*, 106, 1615-1621.

Goffard, N., Frickey, T., and Weiller, G. (2009). PathExpress update: the enzyme neighbourhood method of associating gene-expression data with metabolic pathways. *Nucleic Acids Research*. 37, W335–W339.

Görke, R., Meyer-Bäse, A., Wagner, D., He, H., Emmett, M.R., Conrad, C.A. (2010). Determining and interpreting correlations in lipidomic networks found in glioblastoma cells. *BMC Systems Biology*. 4, 126.

Griffin, J. L., Vidal-Puig, A. (2008). Current challenges in metabolomics for diabetes research: a vital functional genomic tool or just a ploy for gaining funding? *Physiol Genomics,* 34, 1–5.

Hall, R. D. (2005). Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytologist*, 169, 453-468.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–803.

Hollywood, K., Brison, D. R. and Goodacre, R. (2006). Metabolomics: Current technologies and future trends. *Proteomics*, 6, 4716 - 4723.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Holtorf, H., Guitton, M.-C., Reski, R. (2002). Plant functional genomics. *Naturwissenschaften,* 6, 235-249.

Hu, X., Gadbury, G. L., Xiang, Q., Allison, D. B. (2010). Illustrations on using the distribution of a p-value in high dimensional data analysis. *Advances and Applications in Statistical Sciences*, 191-213.

Iba, K., Gibson, S., Nishiuchi, T., Fuse, T., Nishimura, M., Arondel, V., Hugly, S. and Somerville, C. (1993). A gene encoding a chloroplast omega-3 fatty acid desaturase complements alterations in fatty acid desaturation and chloroplast copy number of the Fad7 mutant of *Arabidopsis thaliana*. *The Journal of Biological Chemistry*, 268, 24099-24105.

Klebanov, L., & Yakovlev, A. (2007). How high is the level of technical noise in microarray data? *Biology Direct*.

Kusano, M., Fukushima, A., Arita, M., Jonsson, P., Thomas Moritz, T., Kobayashi, M., Hayashi, N., Tohge, T., Saito, K. (2007). Unbiased characterization of genotype-dependent metabolic regulations by metabolomic approach in *Arabidopsis thaliana*. *BMC Systems Biology*. 1:53.

McLachlan, G. and D. Peel (2000), Finite Mixture Models, Wiley.

Mekhedov, S., de Ilarduya, O.M. and Ohlrogge, J. (2000). Toward a functional catalog of the plant genome. A survey of genes for lipid biosynthesis. *Plant Physiology*, 122, 389-401.

Meng, X.-L. and Rubin, D. B. (1993) Maximum Likelihood Estimation Via the ECM Algorithm: A General Framework, *Biometrika* 80(2): 267-278.

Morgenthal, K., Weckwerth, W., Steuer, R. (2006). Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation. *BioSystems*. 83, 108–117.

Nandi, A., Krothapalli, K., Buseman, C.M., Li, M., Welti, R., Enyedi, A. and Shah, J. (2003). Arabidopsis sfd mutants affect plastidic lipid composition and suppress dwarfing, cell death and the enhanced disease resistance phenotypes resulting from the deficiency of a fatty acid desaturase. *The Plant Cell*, 2383-2398.

Nielsen, J., Oliver, S. (2005). The next wave in metabolome analysis TRENDS in Biotechnology. 23, 544-546.

Okuley, J., Lightner, J., Feldmann, K., Yadav, N., Lark, E. and Browsea, J. (1994). Arabidopsis fad2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *The Plant Cell*, 6, 147-158.

Oliver, S. G. (2002). Functional genomics: lessons from yeast. *Philosophical Transactions of the royal society B*, 357, 17-23.

Oliver, S.G., Winson, M.K., Kell, D.B. and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, 16, 373-378.

Owen, A. B. (2005). Variance of the number of false discoveries. *J. R. Statist. Soc. B*, 411-426.

Oxford Dictionary of Biochemistry, Answers Corporation, *http://www.answers.com/topic/metabolomics*

Oxford Dictionary of Biochemistry, Answers Corporation, *http://www.answers.com/topic/transcriptome#ixzz1Zpq9xu3u*

Paranagama. D. C. (2011). Correlation and variance stabilization in the two group comparison case in high dimensional data under dependencies.

Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., Berden, J. A., Brindle, K. M., Kell, D. B., Rowland, J. J., Westerhoff, H. V., Dam, K. V. and Oliver, S. G. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, 19, 45-50.

Sakurai1, N., Ara, T., Ogata, Y., Sano, R., Ohno, T., Sugiyama, K., Hiruta, A., Yamazaki, K., Yano, K., Aoki, K., Aharoni, A., Hamada, K., Yokoyama, K., Kawamura, S., Otsuka, H.,

Tokimatsu, T., Kanehisa, M., Suzuki, H., Saito, K., Shibata. D. (2011). KaPPA-View4: a metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data. *Nucleic Acids Research.* Database issue, D677-D684.

Schweder T, Spjøtvoll E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika,*69,493-502.

Shulaev, V., Cortes, D., Miller, G., and Mittler, R. (2008). Metabolomics for plant stress response. *Physiologia Plantarum*. 132, 199–208.

Steuer, R. (2006). On the analysis and interpretation of correlations in metabolomic data. B*riefings in Bioinformatics.*7, 151-158.

Steuer, R., Kurths, J., Fiehn, O., Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics*. 19, 1019-1026.

Titterington, D. M. , A. F. M. Smith and U. E. Makov (1985), Statistical Analysis of Finite of Mixture Distributions, John Willey & Son Ltd.

Trethewey. R. N. (2001). Gene discovery via metabolic profiling. *Current opinion in beothechnolog,* 12, 135-138.

U.S. National Library of Medicine's web site: *http://ghr.nlm.nih.gov/handbook/mutationsanddisorders/genemutation*

Urbanczyk-Wochniak, E. and Sumner, L.W. (2007). MedicCyc: a biochemical pathway database for Medicago truncatula. *Bioinformatics*. 23,, 1418–1423.

Wang, X. (2012). Randomization test and correlation effects in high dimensional data. Master report.

Weckwerth, W., Loureiro, M-E, Wenzel, K., and Fiehn, O. (2004). Differential metabolic networks unravel the effects of silent plant phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*. 101, 7809-7814.

Welti, R. and Wang, X. (2004). Lipid species profiling: a high-throughput approach to identify lipid compositional changes and determine the function of genes involved in lipid metabolism and signaling. *Current Opinion in Plant Biology*, 7, 337–344.

Welti, R., Shah, J., Li W., Li, M., Chen, J., Burke, J.J., Fauconnier, M-L, Chapman, K., Chye, M-L, Wang, X. (2007). *Frontiers in Bioscience*, 12, 2494-2506.

Westfall. P. H., Young, S. S. (1993). Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. *1st Edition.* New York,John Wiley & Sons, Inc.

Wirjanto, T. S. and Xu, D. (2009). The Applications of Mixtures of Normal Distributions in Empirical Finance: A Selected Survey. *Working paper*. From website: http://economics.uwaterloo.ca/documents/mn-review-paper-CES.pdf

Wu, L., Winden, W. A. V., Gulik, W. M. V. and Heijnen, J. J. (2005). Application of metabolome data in functional genomics: A conceptual strategy. *Metabolic Engineering.* 7, 302–310.

Zheng, L., Gadbury, G., Shah, J., Welti, R. (2013). Exploration of reactant-product lipid pairs in mutant-wild type lipidomics experiments. Conference Proceedings, *Conference on Applied Statistics in Agriculture*, Kansas State University, Manhattan, KS.

Appendix A - The Lipidomics Experiment Information

Background Information

The modeling plant *Arabidopsis thaliana* are used in this experiment since this plant has small genome and also it is the first plant to be sequenced. The researchers have completed a list of databases for *Arabidopsis thaliana* to search for gene functions or to annotate their own sequence. In this experiment, researchers use *Arabidopsis thaliana* to prove the loss of function of mutated genes when the stress condition is applied to mutant plants compared to wild type plants. The targeted lipidomics data analysis is applied to 9 mutated genes with known gene functions (Fan, 2010). Those mutants include *fad2* (Okuley et al. 1994), *fad3* (Arondel et al. 1992), *fad4* (Gao et al., 2009), *fad5* (Mekhedov et al. 2000), *fad6* (Falcone et al. 1994) and *fad7* (Iba et al. 1993, Gibson et al. 1994). Three other mutants are produced by random mutagenesis. They are *sfd1*, *sfd2* and *sfd3*. The detailed information on the mutants can be found in the paper Nandi et al. (2003). 9 experiments were conducted according to 9 mutants. There are two treatment groups, wild type plants and mutant plants. Since the wild type plant is planted without applying mutation conditions, it will be used as the control group for all 9 different mutants. In each treatment group, 5 samples are randomly chosen and analyzed. The plants were grown by Ashis Nandi who worked with Jyoti Shah from University of North Texas and Christen Buseman who worked with Ruth Welti from the Division of Biology at Kansas State University.

Experiment Material Preparation

Surface sterilized *Arabidopsis thaliana* seeds were germinated on agar plates and then transferred to soil. Plants were grow in a chamber with a 16-h light/8-h dark cycle at 23/21°C under cool fluorescent white light (200 μmol $m^{-2}s^{-1}$) with 58% relative humidity. *Arabidopsis thaliana* (Columbia ecotype) seeds were grown as the wild type plants in the *fad* mutant experiments and *1/8E/5* were used as the wild type plant in the *sfd* mutant experiments. The wild type and the mutant plants were grown in same chamber and sampled in the same time at the same growth stage.

Sample Preparation and Lipid Extraction

Leaves of the plants were harvest and used as the random samples and the lipid extraction procedure are similar with those in the paper Devaiah et al. 2006 and Fan 2010. For completeness, a summary is given below.

- 3 to 5 leaves were quickly immersed in 3 ml isopropanol with 0.01% butylated hydroxytoluene at 75ºC to inactivate lipolytic activity.
- After 15 minutes, 1.5 ml chloroform and 0.6 ml water were added and the tubes were shaken for 1 hour.
- The extract was removed and the leaves were re-extracted with chloroform/methanol (2:1) with 0.01% butylated hydroxytoluene 5 times and each time tubes were shaken for 30 minutes.

The above procedures were stopped until the leaves turned to white to make sure that most of lipids in the leaves were dissolved in the solvent. The remains of the leaf skeleton was dried overnight at 105ºC and weighed to be used for the MS data. The combined extracts were washed once with 1 ml 1 M KCl and once with 2 ml water, the solvent was evaporated under nitrogen, and then the lipid extract was dissolved in 1 ml chloroform.

Mass Spectrometry High-throughput Data Analysis

The targeted lipidomics profiling was performed by using the electrospray tandem mass spectrometry (ESI-MS/MS) (Welti and Wang, 2004). The electrospray tandem mass spectrometry provided high sensitivity comprehensive data analysis for identifying the composition of lipid species. It can identify and quantify the lipid compounds with small amounts of samples. In the experiment, unfractionated lipid extracts were introduced by continuous infusion into the ESI source on a triple quadrupole MS/MS (API 4000, Applied Biosystems, Foster City, CA). Samples were introduced using an autosampler (LC Mini PAL, CTC Analytics AG, Zwingen, Switzerland) fitted with the required injection loop for the acquisition time and presented to the ESI needle at 30 µl/min. More detailed information on the MS data process and the lipid profiling technique can be found in the paper Devaiah et al. 2006.

Appendix B – R Programs

B.1: Correlation Analysis from Fukushima et. al. (2011) in Chapter 2.

```
# Functions:
#     new.names: Get the lipid names in the reduced data. Delete the
#            lipids if the sd = 0.
#     data.s: Get the reduced dataset. Delete the lipids if the
#            sd = 0.
#     datWM: Get the indices of A and B from a matrix (dat) with n1
#            samples from the WT group and n2 samples from the MT
#            group.
#     correl: Calculate spearman's correlation.
#     t.stat: Calculate the single correlation H0: rho = 0.
#     pval.t: Calculate two-tailed p values for the t-test.
#     zstat: Test for the correlation differences H0: rho1-rho2 = 0
#            using Fisher's transformation.
#     p.diff: P values  for the correlation differences.
#     cor.mat: Output the t-test and the Z test results by following
#            the format from additional file 4 in Fukushima et. al.
#            (2011).
##########################################################################

# dat is the input raw data with 141 by 10 matrix. Names is the list
# of lipid names for the raw data.

new.name = function(dat, names){
s = apply(dat,1,sd)
n.names = names[s>0,]
return(n.names)
}

data.s=function(dat){
s = apply(dat, 1, sd)
dat1= dat[s>0,]
return(dat1)
}

datWM = function(dat, n1, n2){
N = n1+n2
n = dim(dat)[1]
w = numeric()
for (i in 1:(n-1)){
     for(j in (i+1):n){
     m1 = c(dat[i,], dat[j,])
     w = c(w, m1, i, j)   # i and j are the indices of A and B lipids.
     }
}
mat = matrix(w, ncol=2*N+2, byrow=TRUE)
return(mat)
```

```
}
correl = function(dat){
n = dim(dat)[1]
corre = c()
p = numeric()
for (i in 1:(n-1)){
           for(j in (i+1):n){
           corre = cor(unlist(dat[i,]), unlist(dat[j,]), method =
           "spearman")
           corre[is.na(corre)] = 0
           p = c(p,corre)
           }
}
return(p)
}

t.stat = function(r, n){
t = r*sqrt((n-2)/(1-r^2))
return(t)
}

zstat=function(r1,r2,n1,n2){
z = (0.5*log((1+r1)/(1-r1))-0.5*log((1+r2)/(1-r2)))/(sqrt(1/(n1-
3)+1/(n2-3)))
return(z)
}

p.diff = 2*pnorm(-abs(zs), lower.tail=TRUE)

cor.mat = function(cor1, cor2, n1, n2){
ts1 = t.stat(cor1, n1)
ts2 = t.stat(cor2, n2)

p1 = pval.t(ts1, df=n1-2)
p2 = pval.t( ts2, df=n2-2)

r.diff = cor1-cor2
zs = zstat(cor1, cor2, n1, n2)
p.diff = 2*pnorm(-abs(zs), lower.tail=TRUE)

library(fdrtool)
fdr.1 = fdrtool(p1, statistic="pvalue")
fdr1 = fdr.1$lfdr
fdr.2 = fdrtool(p2, statistic="pvalue")
fdr2 = fdr.2$lfdr
fdr.diff = fdrtool(p.diff, statistic="pvalue")
fdr.D = fdr.diff$lfdr

mat = data.frame(cor1, p1, cor2, p2, r.diff, p.diff,fdr1, fdr2, fdr.D)
return(mat)
}
```

B.2: Produce Venn Diagram in Figure 2.4.

```
# Function: venn.plot
#      Output: Produce Venn diagram for grouping the correlation pairs
#              in three treatment groups for WT, mtol and tt4 in
#              Figure 2.4. This function can produce the Venn diagram
#              with two events, three events and four events
#              separately in a plot. In this figure, three events are
#              used.
#      list: consist of 2 or 3 or 4 numeric vectors to produce 2 or 3
#            or 4 Venn diagrams.
#      path: the directory for output the Venn diagram.
#      mains: titles for the Venn diagram.
#      filenames: give the file name for the output file.
#########################################################################

# library(VennDiagram)
# library(gridBase)

venn.plot = function(list, path, mains, filenames){
str(list)
lis = lapply(list, lapply, length)
names(lis) = lapply(list, length)
num = length(names(lis))

if(num==2){ # Produce a Venn diagram with 2 events.
venn.diagram(
      x = list,
      filename = paste(path,filenames, ".tiff",sep=""),
            main = mains,
            main.fontface = 2,  # bold face for the main.
            main.cex = 2,
      lwd = 4,
      fill = c("cornflowerblue", "darkorchid1"),
      alpha = 0.75,
      label.col = "white",
      cex = 4,
      fontfamily = "serif",
      fontface = "bold",
      cat.col = c("cornflowerblue", "darkorchid1"),
      cat.cex = 3,
      cat.fontfamily = "serif",
      cat.fontface = "bold",
      cat.dist = c(0.03, 0.03),
      cat.pos = c(-20, 14)
      )
      }

if(num==3){ # Produce a Venn diagram with 3 events.
venn.diagram(
      x = list,
      filename = paste(path,filenames, ".tiff",sep=""),
```

```
            main = mains,
            main.fontface = 2,  # bold face for the main.
            main.cex = 2,
      col = "transparent",
      fill = c("red", "blue", "green"),
      alpha = 0.5,
      label.col = c("darkred", "white", "darkblue", "white", "white",
                  "white", "darkgreen"),
      cex = 2,
      fontfamily = "serif",
      fontface = "bold",
      cat.default.pos = "text",
      cat.col = c("darkred", "darkblue", "darkgreen"),
      cat.cex = 2.5,
      cat.fontfamily = "serif",
      cat.dist = c(0.06, 0.06, 0.03),
      cat.pos = 0
      )
      }

if(num==4){ # Produce a Venn diagram with 4 events.
venn.diagram(
      x = list,
      filename = paste(path, filenames, ".tiff",sep=""),
            main=mains,
            main.fontface=2,  # bold face fo rthe main.
            main.cex = 2,
           col = "black",
      lty = "dotted",
      lwd = 4,
      fill = c("cornflowerblue", "green", "yellow", "darkorchid1"),
      alpha = 0.50,
      label.col = c("orange", "white", "darkorchid4", "white", "white",
                  "white", "white", "white", "darkblue", "white",
                  "white", "white", "white", "darkgreen", "white"),
      cex = 2.5,
      fontfamily = "serif",
      fontface = "bold",
      cat.col = c("darkblue", "darkgreen", "orange", "darkorchid4"),
      cat.cex = 2.5,
      cat.fontfamily = "serif"
      )
      }
}
```

B.3: Produce Test Statistics SSD in Chapter 3.

```
# Function: F.new
#      Output the F ration and the sum square distance SSD. SSD is the
#      Euclidean distance between the WT and MT group centers. The F
#      statistic is the F ratio where the SSD is used in the
#      numerator.
# Input: The raw data with 141 by 10 columns with sample sizes n1 and
#      n2 from WT and MT groups.
#########################################################################

F.new = function( dat, n1, n2){
n = dim(dat)[1]
N = n1+n2
ratio = numeric()
SSD = numeric()

      for(i in 1 : n){
      Aw = dat[i,1: n1]
      Am = dat[i, (n1+1) : N]
      Bw = dat[i, (N+1) : (N+n1)]
      Bm = dat[i, (N+n1+1) : (2*N)]
      A = dat[i, 1 : N]
      B = dat[i, (N+1) : (2*N)]
      dist = sqrt((mean(Aw) - mean(A))^2 + (mean(Bw) - mean(B))^2) +
             sqrt((mean(Am)- mean(A))^2+(mean(Bm) - mean(B))^2)
      MSB = dist^2/1  # numerator of the F_new.
      x1 = cbind(Aw, Bw)
      x2 = cbind(Am, Bm)
      result1 = numeric()
            for (j in 1:n1){
            mat1 = rbind(x1[j,], c(mean(Aw), mean(Bw)))
            d1 = dist(mat1, method='euclidean')
            SSW1 = d1^2
            result1 = c(result1,SSW1)
                       }

      result2 = numeric()
            for (k in 1:n2){
            mat2 = rbind(x2[k,], c(mean(Am), mean(Bm)))
            d2 = dist(mat2, method='euclidean')
            SSW2=d2^2
            result2=c(result2,SSW2)
                       }
      F.new = MSB/((sum(result1) + sum(result2))/(N-2))
      ratio = c(ratio,F.new)
      SSD = c(SSD,dist)  # dis=SSD.
            }

return(cbind(ratio,SSD))  # return F ratio, SSD distance.
}
```

B.4: Generate Bootstrap Under F = G and Make Plots in Chapter 4.

```
# Functions:
#      gofit: Fit the b^th bootstrap samples to Exponential, Gamma,
#             Lognormal and Weibull distributions. Extract the MLEs,
#             AICs and BICs from all model fitting.
#          Input: dat is a bootstrap sample with141 by 10 matrix.
#
#      get.count: Get counts for the minimum KS test statistics D and
#             also the minimum AIC for each bootstrap sample.
#          Input: mat matrix which is part of the output from function
#             gofit. It contains mat[,1:4] to be the minimum
#             statistics D and mat[,5:8] are the minimum AICs from
#             distribution exponential, gamma, lognormal, weibull.
#########################################################################

gofit = function(dat){
library(fitdistrplus)

a = fitdist(dat, "exp")
b = fitdist(dat, "gamma")
d = fitdist(dat, "lnorm")
e = fitdist(dat, "weibull")

rate1 = (a)[[1]]              # rate for exp distribution.
gamm.s = b[[1]][1]
gamm.r = b[[1]][2]
logm = d[[1]][1]              # meanlog
logs = d[[1]][2]              # sdlog
wei.shape = e[[1]][1]         # shape
wei.scale = e[[1]][2]         # scale

f1 = gofstat(a)
f = f1$ks                     # kstest D statistic for exp.
g1 = gofstat(b)
g = g1$ks
h1 = gofstat(d)
h = h1$ks
i1 = gofstat(e)
i = i1$ks

AIC.exp = summary(a)$aic
AIC.gamma = summary(b)$aic
AIC.logn = summary(d)$aic
AIC.wei = summary(e)$aic
BIC.exp = summary(a)$bic
BIC.gamma = summary(b)$bic
BIC.logn = summary(d)$bic
BIC.wei = summary(e)$bic
```

```
return(c(f, g, h, i, AIC.exp, AIC.gamma, AIC.logn, AIC.wei, BIC.exp,
       BIC.gamma, BIC.logn, BIC.wei, rate1, gamm.s, gamm.r, logm,
       logs, wei.shape, wei.scale))
}


get.count = function(mat){
min.D = apply(mat[,1:4], 1, min)
exp.D = ifelse(mat[,1] == min.D, 1, 0)   # 0s or 1s for minimum D.
gamma.D = ifelse(mat[,2] == min.D, 1, 0)
logn.D = ifelse(mat[,3] == min.D, 1, 0)
wei.D = ifelse(mat[,4] == min.D, 1, 0)

Min.AIC = apply(mat[,5:8], 1, min)       # 200 minimum AICs.
exp.A = ifelse(mat[,5] == Min.AIC, 1, 0) # give min.AICs 1.
gamma.A = ifelse(mat[,6] == Min.AIC, 1, 0)
logn.A = ifelse(mat[,7] == Min.AIC, 1, 0)
wei.A = ifelse(mat[,8] == Min.AIC, 1, 0)

exp.a1.count = sum(ifelse(exp.A[which(exp.D == 1)] == 1, 1, 0))
exp.a0.count = sum(ifelse(exp.A[which(exp.D == 1)] == 0, 1, 0))
gam.a1.count = sum(ifelse(gamma.A[which(gamma.D == 1)] == 1, 1, 0))
gam.a0.count = sum(ifelse(gamma.A[which(gamma.D == 1)] == 0, 1, 0))
logn.a1.count = sum(ifelse(logn.A[which(logn.D == 1)] == 1, 1, 0))
logn.a0.count = sum(ifelse(logn.A[which(logn.D == 1)] == 0, 1 ,0))
wei.a1.count = sum(ifelse(wei.A[which(wei.D == 1)] == 1, 1, 0))
wei.a0.count = sum(ifelse(wei.A[which(wei.D == 1)] == 0, 1, 0))

counts = data.frame(cbind(c(exp.a1.count, exp.a0.count),
         c(gam.a1.count, gam.a0.count), c(logn.a1.count,
         logn.a0.count), c(wei.a1.count, wei.a0.count)))
colnames(counts) = c("exponential", "gamma", "lognormal", "Weibull")
rownames(counts) = c("AIC.yes", "AIC.no")

return(counts)
}


#########################################################################
# Bootstrap procedure under the null hypothesis H0: F = G for 9
# datasets. Read in the raw data and output the bootstrap samples into
# the specified file directory.
#########################################################################

dat.path="c:/3.2011research/"
out.path="c:/3.2011research/dissertation/2.fad.boot/"
bio.path="c:/3.2011research/Li.bio.pairs/"
bio.file=list.files("c:/3.2011research/Li.bio.pairs")[1:9]

filenames =c("fad.2_dat", "fad.3_dat", "fad.4_dat", "fad.5_dat",
             "fad.6_dat", "fad.7_dat", "sfd.1_dat", "sfd.2_dat",
             "sfd.3_dat");
```

```
num<-length(filenames);

for(i in 1:num){
name <-read.table("c:/3.2011research/names_141.txt",header=FALSE)
file.names = paste(dat.path,filenames[i],".csv", sep="")
dat<-read.csv(file.names,header=FALSE);
names1=new.name(dat,name)  # function new.name().
red.dat=data.s(dat)  # reduced data

      # bootstrap
      B=200
      for (j in 1:B){
      samp=sample(1:10,10,replace=T)
      dat.b=red.dat[,samp]
      b.names=new.name(dat.b,as.matrix(names1))
      red.dat.b=data.s(dat.b)
      m=t(apply(red.dat.b,1,scale))    # scale reduced data.
      mat=datAB(m)
      A.name=b.names[mat[,21]]   # replace the indexes to A's names.
      B.name=b.names[mat[,22]]   # replace the indexes to B's names.
      AB.names=paste(A.name,B.name, sep="_")

      F.OM = F.new(mat[,1:20]) # output SSD.

      Aw=apply(mat[,1:5],1,mean)
      Am=apply(mat[,6:10],1,mean)
      Bw=apply(mat[,11:15],1,mean)
      Bm=apply(mat[,16:20],1,mean)
      a=Am-Aw
      b=Bw-Bm
      tg=b/a
      y1=ifelse(a>0,1,0)
      y2=ifelse(b>0,1,0)
      y=y1+y2

      scaled=data.frame(mat,A.name,B.name,AB.names,y,a,b,tg,F.OM[,1],
                        F.OM[,2])
      names(scaled)=c("Aw1", "Aw2", "Aw3", "Aw4", "Aw5", "Am1", "Am2",
                      "Am3", "Am4", "Am5", "Bw1", "Bw2", "Bw3", "Bw4",
                      "Bw5", "Bm1", "Bm2", Bm3", "Bm4", "Bm5",
                      "A_index", "B_index", "A.name", "B.name",
                      "AB.name", "y", "avg(Am)-avg(Aw)",
                      "avg(Bw)-avg(Bm)", "tg", "F_OM", "SSD")

      # extract the elements with y = 2.

      y.scaled=scaled[scaled[,26]==2,]
      write.csv(y.scaled, file=paste(out.path,'All.metric.y.dat.',
                filenames[i],j,'.boot.csv', sep=''),row.names=FALSE)
      }
}
```

```
#########################################################################
# Read in the bootstrap samples under the null hypothesis H0: F = G for
# fad2 dataset. Overlaid the chosen parametric distribution with the
# empirical bootstrap distribution.
#########################################################################

# Functions:
#      lim: input the bootstrap statistics matrix.
#           Output: The 5th, mean, and 95th percentile parametric null
#                   distribution parameters.
# ------------------------------------------------------------------------

lim=function(mat){
means = round(apply(mat, 2, mean),2)
sds=apply(mat,2,sd)

LL=round(means-1.96*sds/sqrt(200),2)
UL=round(means+1.96*sds/sqrt(200),2)

quant=function (x) {quantile(x,  probs = c(0.05, 0.95))}
fifth=t(apply(mat,2,quant))

limits=data.frame(LL,UL,means,fifth,row.names=c("exp.rate",
                  "gam.shape,"gam.rate", "log.mean", "log.sd",
                  "wei.shape", "wei.scale"))
colnames(limits)=c("LL","UL","means","5th","95th")

return(limits)
}


#--------------------------------
# Read in 200 bootstrap samples.
#--------------------------------

b.F=list.files()[1:200]
len=length(b.F)

# Read in the actual fad2 dataset.
fad2=read.csv("c:/3.2011research/All.metric.y.dat.fad.2_dat.csv")
colnames(fad2)
SSD.fad2=fad2[,36]
tg.fad2=fad2[,34]

R.fad2=(SSD.fad2-2.684)^2+(tg.fad2-1)^2
nlogR.fad2= -log(R.fad2)
s.nlogR.fad2 = nlogR.fad2+5   # adding a shift by +5.

tg.ds = matrix(nrow=200, ncol=19,byrow=T)
SSD.ds = matrix(nrow=200, ncol=19,byrow=T)
R.ds = matrix(nrow=200, ncol=19,byrow=T)
```

```
# shifted negative logR matrix.
s.nlogR.ds = matrix(nrow=200, ncol=19,byrow=T)

big.tg = numeric()
big.SSD = numeric()
big.R = numeric()
big.s.nlogR=numeric()
count=numeric()

# quantile of the shifted nlogR.
q.tg.ds = matrix(nrow=200, ncol=100,byrow=T)
q.SSD.ds = matrix(nrow=200, ncol=100,byrow=T)
q.s.nlogR.ds = matrix(nrow=200, ncol=100,byrow=T)

for (i in 1:len) {
u<-read.csv(b.F[i])
u=u[u[,29]<=10,]
tg.boot=c(u[,29])
SSD.boot=c(u[,31])

R.boot = (SSD.boot-2.684)^2 + (tg.boot-1)^2
logR = log(R.boot)
s.nlogR.boot = -log(R.boot)+5
tg.ds[i,] = gofit(tg.boot)
SSD.ds[i,] = gofit(SSD.boot)
s.nlogR.ds[i,] = gofit(s.nlogR.boot)

q.tg.ds[i,]=quantile(tg.boot, c(seq(0.01,1,by = 0.01)))
q.SSD.ds[i,]=quantile(SSD.boot, c(seq(0.01,1,by = 0.01)))
q.s.nlogR.ds[i,]=quantile(s.nlogR.boot, c(seq(0.01,1,by = 0.01)))

count=c(count,length(tg.boot))
big.tg = c(big.tg, tg.boot)
big.SSD = c(big.SSD, SSD.boot)
big.R = c(big.R, R.boot)
big.s.nlogR = c(big.s.nlogR,s.nlogR.boot)
}

colnames(SSD.ds)= c("exp.D", "gamma.D", "lognorm.D", "Weibull.D",
            "AIC.exp", "AIC.gamma", "AIC.logn", "AIC.wei", "BIC.exp",
            "BIC.gamma", "BIC.logn", "BIC.wei", "rate1", "gamm.s",
            "gamm.r", "logm", "logs", "wei.shape", "wei.scale" )

colnames(tg.ds)= c("exp.D", "gamma.D", "lognorm.D", "Weibull.D",
           "AIC.exp", "AIC.gamma", "AIC.logn", "AIC.wei", "BIC.exp",
           "BIC.gamma", "BIC.logn", "BIC.wei", "rate1", "gamm.s",
           "gamm.r", "logm", "logs", "wei.shape", "wei.scale" )

colnames(s.nlogR.ds)= c("exp.D", "gamma.D", "lognorm.D", "Weibull.D",
          "AIC.exp", "AIC.gamma", "AIC.logn", "AIC.wei", "BIC.exp",
          "BIC.gamma", "BIC.logn", "BIC.wei", "rate1", "gamm.s",
          "gamm.r", "logm", "logs", "wei.shape", "wei.scale" )
```

```
#-----------------------------------------------------------------------
# Produce Figure 4.4: box plots for the K-S statistics D distribution
#-----------------------------------------------------------------------

par(mfrow=c(1,3))
boxplot(tg.ds[,1:4], xlab = " ", ylab = "ks.test statistic D",
        main = expression('statistic D distribution for tg'^{*}))

boxplot(ssd.ds[,1:4], xlab = " ", ylab = "ks.test statistic D",
        main = expression('statistic D distribution for SSD'^{*}))

boxplot(s.nlogR.ds[,1:4], xlab = " ", ylab = "ks.test statistic D",
        main = expression('statistic D distribution for R'[T]))


#---------------------------
# Produce Figure 4.5 for SSD
#---------------------------

library("Rlab")
y=c(seq(0.01,1,by = 0.01))

quant1=function(x){quantile(x,c(0.05))}
quant2=function(x){quantile(x,c(0.95))}
quant3=function(x){quantile(x,c(0.25))}
quant4=function(x){quantile(x,c(0.75))}

## Figure 4.5 (a).
plot(apply(q.SSD.ds, 2, median), y, type="l", lwd=3, col="red",
     xlab="SSD.boot", ylab="Cumulative probability", main="ECDF and
     CDFs comparison for SSD.boot")

lines(apply(q.SSD.ds, 2, quant3), y, lty=1, lwd=1, col="red")
lines(apply(q.SSD.ds, 2, quant4), y, lty=1, lwd=1, col="red")

plot(function(x) pexp(x, rate = lim.SSD[1, 3]),
     from = min(big.SSD), to = max(big.SSD),
     add = TRUE, lty = 2,lwd = 2,col="blue")

plot(function(x) pgamma(x, shape = lim.SSD[2,3] ,
     rate = lim.SSD[3, 3]),
     from = min(big.SSD), to = max(big.SSD),
     add = TRUE, lty = 2,lwd = 2,col = "purple")

plot(function(x) plnorm(x, meanlog = lim.SSD[4, 3],
     sdlog = lim.SSD[5, 3]),
     from = min(big.SSD), to = max(big.SSD),
     add = TRUE, lty = 2,lwd = 2)

plot(function(x) pweibull(x, shape = lim.SSD[6, 3],
     scale = lim.SSD[7, 3]),
```

```
     from = min(big.SSD), to = max(big.SSD),
     add = TRUE, lty = 5,lwd = 2,col = "green")

legend("bottomright", legend = c("ECDF", "ECDF-25%,75%tiles",
     "CDF-exp", "CDF-weibull", "CDF-gamma", "CDF-lognormal"),
     lwd = c(2,1,2,2,2,2), col = c("red", "red", "blue", "green",
     "purple", "black"), lty = c(1,1,2,5,2,2))

## Figure 4.5 (b).
plot(apply(q.SSD.ds, 2, median), y, type="l", lwd = 3, col = "red",
     xlab = "SSD.boot", ylab = "Cumulative probability",
     main= "ECDF and CDFs comparison for SSD.boot")

lines(apply(q.SSD.ds, 2, quant1), y, lty = 1, lwd = 1, col = "red")
lines(apply(q.SSD.ds, 2, quant2), y, lty = 1, lwd = 1, col = "red")

Pos = seq(0.01, 1, by = 0.05)
bplot(q.SSD.ds[,pos*100], pos = pos, label.cex = 0,
      horizontal = TRUE, add = TRUE)

legend("bottomright", legend = c("CDF-Webull"), lwd = c(3, NA),
      col = c("blue", NA),lty = c(1, NA))


#-------------------------------------------------
# Produce Figure 4.8 for overlaid R_T distribution
#-------------------------------------------------

p2.95 = apply(q.s.nlogR.ds, 2, quant2)
hist(s.nlogR.fad2, ylim = c(0, 0.7), prob = TRUE,
     xlab = expression('R'[T]), col = "green", density = 55,
     main = expression('Distributions overlaid for R'[T]))

hist(p2.95, breaks = 35, prob = TRUE, add = TRUE,
     col = "pink", density=60)

plot( function(x) dweibull(x, shape = lim.s.nlogR[6,5],
    scale = lim.s.nlogR[7,5]), from = 0, to = 13, add = TRUE,
    lty = 2,lwd = 2,col = "blue")

legend("topright", legend = c("95th Weibull null", "95th empirical",
    expression('fad2 R'[T])), col = c("blue", "pink", "green"),
    lty = c(2, NA, NA),lwd = c(2, NA, NA),fill = c(NA, "pink",
    "green"), border = FALSE)
```

B.5: Mixture Normal Distributions in Chapter 5.

```
# Functions:
#     m.norm: find the values of (1 – CDF) of the mixture normal with
#             2 components. Inputs are statistics x, mu, sigma and the
#             mixing proportions. Output the p values for each
#             statistic.
#     plot.mix: Produce the normal mixture plot with 2 normal curves
#             and one mixture curve. Inputs are a set of parameters
#             and the number of component. Output will a figure.
#########################################################################

m.norm <- function(x, mu, sigma, prop) {
ps = prop[1] * pnorm(x, mu[1], sigma[1]) + (1 - prop[1]) * pnorm(x,
mu[2], sigma[2])
return(1-ps)
}

plot.mix=function(par, compnum){
     mix=matrix(par, ncol=compnum, nrow=3, byrow=TRUE)
     for(i in 1:compnum){
     curve(mix[1,i]*dnorm(x,mean=mix[2,i],sd=mix[3,i]), add=TRUE)
     }
}


## Produce Figure 5.2.

library(mixtools)
boot.nlogR = c(big.s.nlogR, s.nlogR.fad4)

par(mfrow=c(1,3))
hist(s.nlogR.fad4, prob = TRUE, ylim = c(0, 0.6), breaks = 35,
    col = "green", density = 45, xlab = expression('R'[T]),
    main = expression('(a). The Distribution of R'[T]),
    font.main = 2, cex.lab = 1.5, cex.axis = 1, cex.main = 1.8)

hist(boot.nlogR, prob = TRUE, col = "orange", breaks=35, density=90,
    ylim = c(0, 0.7), xlab = expression('R'[T]^{paste("*")}),
    main = expression('(b). The Distribution of R'[T]^{paste("*")}),
    cex.lab = 1.5, cex.axis = 1, cex.main = 1.8)

hist(s.nlogR.fad4, prob = TRUE, ylim = c(0, 0.8), breaks = 35,
    col = "green", density = 45, xlab = expression('R'[T]),
    main="(c). Distribution Overlaid", font.main = 2, cex.lab = 1.5,
    cex.axis = 1, cex.main = 1.8)

hist(boot.nlogR, prob = TRUE, col = "orange", breaks = 35,
    density = 90, ylim = c(0, 0.7), add = TRUE, main="")

mix.big2 = normalmixEM(boot.nlogR,k = 2, maxit=1000,epsilon=0.01)
mix.big3 = normalmixEM(boot.nlogR,k = 3, maxit=1000,epsilon=0.01)
```

```
pars.2=c(mix.big2$lambda,mix.big2$mu,mix.big2$sigma)
pars.3=c(mix.big3$lambda,mix.big3$mu,mix.big3$sigma)


# Produce Figure 5.3: Fit two-component mixture normal distribution.

hist(s.nlogR.fad4, prob = TRUE, ylim = c(0, 0.8), breaks = 35,
     col = "green", density = 45, xlab = expression('R'[T]),
     main = "Two Components Mixture Normal")

hist(boot.nlogR, prob = TRUE, col = "orange", breaks = 35,
     density = 90, ylim = c(0, 0.7), add = TRUE, main="")

plot.mix(pars.2, 2)
x<-seq(min(boot.nlogR), max(boot.nlogR),.01)
Fx1 <- pars.2[1] * dnorm(x, pars.2[3], pars.2[5])
Fx2 <- pars.2[2] * dnorm(x, pars.2[4], pars.2[6])
Fx = Fx1 + Fx2
lines(x,Fx,lty=2, lwd=2, col="red")

legend("topright", legend = c("components","Mixture"),
     lty = c(1, 2),lwd = c(2, 2),col = c("black", "red"),
     border = FALSE)


# Produce Figure 5.4: Fit three-component mixture normal distribution.

hist(s.nlogR.fad4, prob = TRUE, ylim = c(0, 0.8), breaks = 35,
     col = "green", density = 45, xlab = expression('R'[T]),
     main = "Three Components Mixture Normal")

hist(boot.nlogR, prob = TRUE, col = "orange", breaks = 35,
     density = 90, ylim = c(0, 0.7), add = TRUE, main="")
     plot.mix(pars.3, 3)
     Fx1 <- pars.3[1] * dnorm(x, pars.3[4], pars.3[7])
     Fx2 <- pars.3[2] * dnorm(x, pars.3[5], pars.3[8])
     Fx3 <- pars.3[3] * dnorm(x, pars.3[6], pars.3[9])
     Fx = Fx1 + Fx2 + Fx3
     lines(x, Fx, lty = 2, lwd = 2, col = "red")

legend("topright", legend = c("components", "Mixture"), lty = c(1,2),
     lwd = c(2, 2), col = c("black", "red"), border = FALSE)


## Produce Table 5.3: the 95% CI for the parameters for three-
## component mixture.

summary(mix.big3)
ses2 = boot.se(mix.big3, B = 100, arbvar = FALSE)
ses2$lambda.se
ses2$mu.se
ses2$sigma.se
```

```
para2 = c(mix.big3$lambda, mix.big3$mu, mix.big3$sigma)
se2 = c(ses2$lambda.se, ses2$mu.se, ses2$sigma.se)

# z interval.

mt4 = matrix(nrow=9, ncol = 2, byrow = TRUE)
for (i in 1:length(para2)){
     mt4[i,] = round(para2[i] + c(-1, 1) * crit.val * se2[i],3)
}

mt5 = cbind(para2, se2, mt4)
colnames(mt5)=c("Estimate", "SE", "LL", "UL")
row.names(mt5)=c("prop1", "prop2", "prop3", "mu1", "mu2", "mu3",
                 "sigma1", "sigma2", "sigma3")
mt5   # Table 5.3
```

B.6: Mixture Normal Distributions to Fit the Empirical Distribution of R_T in Chapter 7.

```
# Test the number of components in a normal mixture model.
# Step 1: fit the 2, 3 and 4 components mixture 100 times to pick out
#         the maximum loglik from 100 model fits.
# Step 2: test Two-component vs. Three-component
# Step 3: test Three-component vs. Four-component
#########################################################################

# Step1.

mat2 = matrix(, nrow = 100, ncol = 7, byrow = TRUE)
mat3 = matrix(, nrow = 100, ncol = 10, byrow = TRUE)
mat4=matrix(, nrow = 100, ncol = 13, byrow = TRUE)

library(mixtools)
for (j in 1:100){
mix.R2 = normalmixEM(nlogR, k = 2, maxit = 1000, epsilon = 0.01)
mix.R3 = normalmixEM(nlogR, k = 3, maxit = 1000, epsilon = 0.01)
mix.R4 = normalmixEM(nlogR, k = 4, maxit = 1000, epsilon = 0.01)

mat2[j,] = c(unlist(mix.R2[c("lambda", "mu", "sigma")]),mix.R2$loglik)
mat3[j,] = c(unlist(mix.R3[c("lambda", "mu", "sigma")]),mix.R3$loglik)
mat4[j,] = c(unlist(mix.R4[c("lambda", "mu", "sigma")]),mix.R4$loglik)
}

par2 = as.vector(mat2[which.max(mat2[,7]),])
par3 = as.vector(mat3[which.max(mat3[,10]),])
par4 = as.vector(mat4[which.max(mat4[,13]),])

mat.comp2[i,] = par2
mat.comp3[i,] = par3
```

```
mat.comp4[i,] = par4

loglik2 = par2[7]
loglik3 = par3[10]
loglik4 = par4[13]

stat2 = loglik3 - loglik2  # observed statistic for step2.
stat3 = loglik4 - loglik3  # observed statistic for step3.

## Step2: testing the number of components: Two-component vs. Three-
## components.

p = par2[1];
mu1 = par2[3];
mu2 = par2[4];
sig1 = par2[5];
sig2 = par2[6]

len = length(nlogR)
stat.s = numeric()

for(k in 1:1000){
samp = runif(len)
cont.f1 = sum(ifelse(samp < p, 1, 0))

nlogR.f1 = rnorm(cont.f1, mu1, sig1)
nlogR.f2 = rnorm(len-cont.f1, mu2, sig2)

logR.star = c(nlogR.f1, nlogR.f2)

mix2s = normalmixEM(logR.star, k = 2, maxit = 1000, epsilon = 0.01)
mix3s = normalmixEM(logR.star, k = 3, maxit = 1000, epsilon = 0.01)

star = mix3s$loglik - mix2s$loglik
stat.s = c(stat.s, star)   # stat stars.
}

pval2 = mean(stat.s > stat2) # p-value = 0.038


## Step3: testing the number of components: Three-component vs. Four-
## components.

p1 = par3[1]; p2 = par3[2]; p3 = par3[3]
mu1 = par3[4]; mu2 = par3[5]; mu3 = par3[6]
sig1 = par3[7]; sig2 = par3[8]; sig3 = par3[9]

stat3.s = numeric()

for(l in 1:1000){
samp = runif(len)
cont.f1 = sum(ifelse(samp < p1, 1, 0))
```

```
cont.f3 = sum(ifelse(samp >= p1+p2, 1, 0))
cont.f2 = len-cont.f1-cont.f3

nlogR.f1 = rnorm(cont.f1, mu1, sig1)
nlogR.f2 = rnorm(cont.f2, mu2, sig2)
nlogR.f3 = rnorm(cont.f2, mu3, sig3)

logR.stars = c(nlogR.f1, nlogR.f2, nlogR.f3)

mix3s = normalmixEM(logR.stars, k = 3, maxit = 1000,epsilon = 0.01)
mix4s = normalmixEM(logR.stars,k = 4, maxit = 1000, epsilon = 0.01)

stars = mix4s$loglik - mix3s$loglik
stat3.s = c(stat3.s, stars)   # stat stars.
}

pval3 = mean(stat3.s > stat3) # pval = 0.184


##---------------------------------------------------------------------
## Find the posterior probabilities using three-component mixture
## model. Output all the results. Make posterior probability plot.
##---------------------------------------------------------------------

library(mixtools)
attributes(mix.R3)
post = mix.R3$posterior
post
dt = cbind(nlogR.fad2, post[,3])

# get the posterior probabilities for the biologically pairs.

fad2$prob3 = post[,3]
colnames(fad2)

bio.fad2 = read.table("c:/bio.nams_fad.2.txt")
colnames(fad2)
bio.dat = fad2[fad2$AB.name %in% bio.fad2[,1],]
bio.tg = bio.dat$tg
bio.SSD = bio.dat$SSD
bio.nlogR = bio.dat$nlogR
bio.dat = bio.dat[with(bio.dat, order(-nlogR)),]
bio.prob3 = bio.dat$prob3

prop = numeric()
num.pairs = numeric()
num.distin = numeric()
dist.list = list()

for (i in 1:length(bio.prob3)){
proportion = mean(fad2$prob3>bio.prob3[i])
prop = c(prop, proportion)
```

```
num.pairs = c(num.pairs, proportion*4623)

react = fad2[fad2$prob3 > bio.prob3[i],]$A.name
distin.react = unique(react)

num.distin = c(num.distin, length(distin.react))

write.csv(distin.react,
         file = paste("c:/dissertation/
         A.three.comp", i, ".csv", sep=""), row.names = FALSE)

dist.list = list(distin.react[i])
}

Prop
num.pairs # number of significant pairs for total 4623 y = 2 file.
num.distin
dist.list

mat = data.frame(bio.dat[,25], bio.dat$nlogR, bio.prob3, prop,
      num.pairs, num.distin)

colnames(mat)=c("AB.name", "nlogR", "posterior.prob", "Proportion",
    "number.sig.pairs", "num.distinct.react")

write.csv(mat,
    file="c:/dissertation/posterior.three.comp.csv",row.names=FALSE)


# Poster probability plots in Figure 7.5 (a).

plot(nlogR.fad2, post[,3], type = "n", xlab = expression('R'[T]),
     ylab = "Posterior Probability",main = "Posterior probabilities")
     mat1 = dt[order(dt[, 1.]), ]
     lines(mat1[, 1.], mat1[, 2.])

abline(v = bio.nlogR, col = "red", lty = 2, lwd = 2.6)

points(bio.nlogR, bio.prob3)

legend("topleft", legend = c("Bio-feasible lipid pair"),
       lty = c(2),lwd = c(2.6), col = c("red"))
```

B.7: Simulation in Chapter 9.

```
#-----------------------------------------------------------------------
# Step 1: Calculate mean and sd vectors in the actual data fad4.
#-----------------------------------------------------------------------

name <- read.table("c:/names_141.txt",header=FALSE)
dat <- read.csv("c:/fad.4_dat.csv",header=FALSE)
names1 = new.name(dat, name)  # function new.name().
dt = data.s(dat)              # reduced data
dim(dat)
dim(dt)                       # 129 by 10.

wt1=dt[,1:5]
mt1=dt[,6:10]


# Lipids with indices 12, 60, 102 have sd = 0 in wt.
wt.m1=as.vector(apply(wt1,1,mean))
wt.s1=as.vector(apply(wt1,1,sd))

length(wt.s1)
length(names1)


# Lipids with indices 14, 24, 25, 90, 116, 124 have sd = 0.
mt.m1 = as.vector(apply(mt1, 1, mean))
mt.s1 = as.vector(apply(mt1, 1, sd))

# Delete the lipid if the WT or MT sd = 0.
not = c(which(wt.s1 == 0), which(mt.s1 == 0))
fad4 = dt[-not,]
dim(fad4)   # 120 by 10

lip.n = names1[-not]  # lipid names.
length(lip.n)

wt = fad4[,1:5]   # 120 by 5
mt = fad4[,6:10]   # 120 by 5

# Mean and sd vectors in the WT and MT groups.
wt.m = as.vector(apply(wt, 1, mean))
wt.s = as.vector(apply(wt, 1, sd))
mt.m = as.vector(apply(mt, 1, mean))
mt.s = as.vector(apply(mt, 1, sd))
```

```
#-----------------------------------------------------------------
# Step 2: Simulate realistic data that is close to the actual data.
#-----------------------------------------------------------------

# Simulate correlation matrix R_W in (9.1), here cor is R_W.

r=diag(30)
r[which(r==0)]=0.5
r

library(Matrix)
r1=bdiag(r, diag(dim(wt)[1]-30))
r1
cor=as.matrix(r1)
cor

# Simulate correlation matrix R_M in (9.2), here cor1 is R_M with the
first 7 pairs to be negative correlations.

cor1=cor
cor1[1,2:8] = -0.5
cor1[2:8,1] = -0.5
cor1
## Simulate realistic data.

library(MBESS)
library(Matrix)

sigma.wt1 = cor2cov(cor, wt.s) # covariance for WT.
sigma.mt1 = cor2cov(cor1, mt.s) # covariance for MT is not positive
                                  definite.
# change the covariance in MT to be positive definite.
Sigma = nearPD(sigma.mt1, corr = FALSE,
               keepDiag = FALSE, do2eigen = TRUE)

sigma.m = as.matrix(sigma$mat)  # Covariance matrix for the WT.
dim(sigma.m)# 120 by 120

## generate multivariate normal data with sample size 5.

library(mvtnorm)
x.w1 = t(rmvnorm(5, wt.m, sigma.wt1))
x.m1 = t(rmvnorm(5, mt.m, sigma.m))
sim1 = cbind(x.w1, x.m1)
sim1
```

```
#-------------------------------------------------------------------
# Step 3: Simulate null data using the common correlation matrix
#         corr = R_W.
#-------------------------------------------------------------------

r = diag(30)
r[which(r == 0)] = 0.5
r

library(Matrix)
r1 = bdiag(r, diag(dim(wt)[1]-30))
r1

cor=as.matrix(r1)
cor

library(MBESS)
sigma.wt=cor2cov(cor,wt.s)      # covariance for WT.
sigma.mt=cor2cov(cor,mt.s)      # covariance for WT.
dim(sigma.mt)                   # 120 by 120

## generate multivariate normal data with sample size 5..
library(mvtnorm)
x.w = t(rmvnorm(5, wt.m, sigma.wt))
x.m = t(rmvnorm(5, mt.m, sigma.mt))
sim.null = cbind(x.w, x.m)
sim.null


#-------------------------------------------------------------------
# Step 4: Generate bootstrap samples using the null data under the
#         null hypothesis  μ_F = μ_G .
#-------------------------------------------------------------------

out.path = "c:/simulation/boot.n5/"
n.names = new.name(sim.null, as.matrix(lip.n))  # function new.name().
r.dat = data.s(sim.null)

dim(sim.null)
length(lip.n)
length(n.names)
dim(r.dat)

# Remove the group means and center the overall mean to the overall
# mean.

r.mean.wt = apply(r.dat[,1:5], 1, mean)
r.mean.mt = apply(r.dat[,6:10], 1, mean)
o.mean = apply(r.dat, 1, mean)
mod.WT = r.dat[,1:5] - r.mean.wt + o.mean
mod.MT = r.dat[,6:10] - r.mean.mt + o.mean
```

```
mod.data = cbind(mod.WT, mod.MT)

# bootstrap samples.

B=200
for (j in 1:B){
samp.w = sample(1:5, 5, replace = T)  # resample within WT and MT.
samp.m = sample(1:5, 5, replace = T)

dat.b = cbind(mod.WT[ ,samp.w], mod.MT[ ,samp.m])
b.names = new.name(dat.b, as.matrix(n.names))
dat.b1 = data.s(dat.b)

m = t(apply(dat.b1, 1, scale)) # scaled data.
Mat = datAB(m)

A.name = b.names[mat[ ,21]]   # replace the indexes to A's names.
B.name = b.names[mat[ ,22]]   # replace the indexes to B's names.
AB.names = paste(A.name, B.name, sep="_")

F.OM = F.new(mat[ ,1:20])      # output the F.OM ratio and SSD.

# Find y = 2 file.
Aw = apply(mat[ ,1:5], 1, mean)
Am = apply(mat[ ,6:10], 1, mean)
Bw = apply(mat[ ,11:15], 1, mean)
Bm = apply(mat[ ,16:20], 1, mean)
a = Am-Aw
b = Bw-Bm
tg = b/a
y1 = ifelse(a>0, 1, 0)
y2 = ifelse(b>0, 1, 0)
y = y1 + y2

Scaled = data.frame(mat, A.name, B.name, AB.names,
                    y, a, b, tg, F.OM[ ,1], F.OM[ ,2])

names(scaled) = c("Aw1", "Aw2", "Aw3", "Aw4", "Aw5", "Am1", "Am2",
                  "Am3", "Am4", "Am5", "Bw1", "Bw2", "Bw3", "Bw4",
                  "Bw5", "Bm1", "Bm2", "Bm3", "Bm4", "Bm5", "A_index",
                  "B_index", "A.name", "B.name", "AB.name", "y",
                  "avg(Am)-avg(Aw)", "avg(Bw)-avg(Bm)", "tg", "F_OM",
                  "SSD")

# extract the elements with y = 2.
y.scaled = scaled[scaled$y == 2,]

write.csv(null.scale, file = paste(out.path, 'sim.null.n5.', j,
         '.boot.csv', sep=''), row.names = FALSE)
}
```